DOE/NV/13609-27

A Validation Process for the Groundwater Flow and Transport Model of the Faultless Nuclear Test at Central Nevada Test Area

Prepared by

Ahmed Hassan

submitted to

Nevada Site Office
National Nuclear Security Administration
U.S. Department of Energy
Las Vegas, Nevada

JANUARY 2003

Publication No. 45197

DOE/NV/13609-24

A Validation for the Groundwater Flow and Transport Model of the Faultless Nuclear Test at Central Nevada Test Area

Prepared by

Ahmed Hassan
Division of Hydrologic Sciences
Desert Research Institute
University and Community College System of Nevada

Publication No. 45197

Submitted to

Nevada Site Office
National Nuclear Security Administration
U.S. Department of Energy
Las Vegas, Nevada

January 2003

The work upon which this report is based was supported by the U.S. Department of Energy under Contract #DE-AC08-00NV13609. Approved for public release; further dissemination unlimited.

EXECUTIVE SUMMARY

Many sites of groundwater contamination rely heavily on complex numerical models of flow and transport to develop closure plans. This has created a need for tools and approaches that can be used to build confidence in model predictions and make it apparent to regulators, policy makers, and the public that these models are sufficient for decision making. This confidence building is a long-term iterative process and it is this process that should be termed "model validation." Model validation is a process not an end result. That is, the process of model validation cannot always assure acceptable prediction or quality of the model. Rather, it provides safeguard against faulty models or inadequately developed and tested models. Therefore, development of a systematic approach for evaluating and validating subsurface predictive models and guiding field activities for data collection and long-term monitoring is strongly needed. This report presents a review of model validation studies that pertain to groundwater flow and transport modeling. Definitions, literature debates, previously proposed validation strategies, and conferences and symposia that focused on subsurface model validation are reviewed and discussed. The review is general in nature, but the focus of the discussion is on site-specific, predictive groundwater models that are used for making decisions regarding remediation activities and site closure. An attempt is made to compile most of the published studies on groundwater model validation and assemble what has been proposed or used for validating subsurface models. The aim is to provide a reasonable starting point to aid the development of the validation plan for the groundwater flow and transport model of the Faultless nuclear test conducted at the Central Nevada Test Area (CNTA).

The review of previous studies on model validation shows that there does not exist a set of specific procedures and tests that can be easily adapted and applied to determine the validity of site-specific groundwater models. This is true for both deterministic and stochastic models, with the latter posing a more difficult and challenging problem when it comes to validation. This report then proposes a general validation approach for the CNTA model, which addresses some of the important issues recognized in previous validation studies, conferences, and symposia as crucial to the process. The proposed approach links model building, model calibration, model predictions, data collection, model evaluations, and model validation in an iterative loop. The approach focuses on use of collected validation data to reduce model uncertainty and narrow the range of possible outcomes of stochastic numerical models. It accounts for the stochastic nature of the numerical CNTA model, which used Monte Carlo simulation approach. The proposed methodology relies on the premise that absolute validity is not even a theoretical possibility and is not a regulatory requirement. Rather, it highlights the importance of testing as many aspects of the model as possible and using as many diverse statistical tools as possible for rigorous checking and confidence building in the model and its predictions. It is this confidence that will eventually allow for regulator and public acceptance of decisions based on the model predictions.

CONTENTS

LIST OF FIGURES

1. INTRODUCTION

Many of the most difficult environmental problems facing scientists pertain to groundwater contamination. Contaminants range from solvents and heavy metals to radionuclides, and they come from a wide variety of sources ranging from leaking tanks to underground nuclear tests. All of these problems have in common the need to demonstrate understanding of past, present, and future migration behavior in subsurface systems where there are limited opportunities to make observations. As a result, resolution of groundwater contamination problems relies extensively on numerical models of flow and transport. Great advances in both theoretical and applied areas of numerical modeling have been made in recent years, driven in large part by advances in computer resources. This has enabled sophisticated incorporation of the uncertainty inherent in all analyses of the subsurface through use of stochastic techniques (e.g., Dagan, 1989; Gelhar, 1993; Cushman, 1997).

During the past two decades, hydrogeologic studies have commonly used stochastic tools to incorporate the effects of spatial variability of hydrologic properties and parametric uncertainty into the predictive capabilities of numerical groundwater flow and contaminant transport models. These studies have made it clear that inadequate and insufficient data limit the ability of these models to predict system behavior without substantial uncertainty (e.g., Pohll *et al.*, 1999; Pohlmann *et al.*, 2000; Hassan *et al.*, 2001). Uncertainty is always inherent in the model prediction and is the result of the inability to fully characterize the subsurface environment and the processes controlling the system behavior. Full characterization is limited by access to the subsurface, which requires extensive borehole drilling that can adversely affect the integrity of the geologic structure of the site or be prohibitively expensive. The stochastic tools used to overcome or address the issues of uncertainty provide predictions as ranges of output with associated probabilities or confidence levels.

The significant advances in computational resources made in the past decade have elevated the level of complexity of numerical and analytical stochastic models to such a high level that a gap has been created between model results and confident assessment of the accuracy (or at least relevance) of model simulations by regulators and the public. The acceptance of the model results by the regulators and the public is an essential prerequisite to close subsurface contaminated sites. This acceptance is difficult to attain with the large range of uncertainty associated with the predictions of these stochastic models. Inclusion of a model validation phase is probably the best way to address this problem and it can achieve buy-in for a closure process involving numerical groundwater modeling. Model validation is the process of evaluating and testing the different aspects of the model for the purpose of refining, enhancing, and building confidence in the model predictions in such a way that allows for sound decision-making. It is the process that follows the determination that the model is well developed and calibrated, after sensitivity analysis indicates insignificant uncertainty reduction from additional characterization efforts. At this stage, and to allow for making decisions based on the model results, the model validation process should start. Model validation is thus a process, not an end result by itself. It cannot ensure an acceptable model. Rather, it provides a safeguard against faulty models or inadequately developed and tested models. If the validation process indicates that major deficiencies exist in the model and a new round of characterization, conceptualization, calibration, modeling, and prediction is needed, it does not mean that the validation process failed. On the contrary, this means that the "process" is successful in achieving its objectives. If the refined model results are proven (through the validation process) to be not in any major

contradiction with field data and these results end up being used as the basis for decision-making, then the validation process indicates that the model is valid for making decisions (not necessarily a true or exact representation of reality).

Regulators and decision makers should understand that there is no way to guarantee that a model-based decision is always correct, or that a model can ever be proven to be valid in the ***strictest sense*** of the term (van der Heijde, 1990). Many assert that it is impossible to validate a groundwater numerical model because such a claim would assert a demonstration of truth that can never be attained for our approximate solutions to subsurface problems (Oreskes *et al.*, 1994). These views consider the validation from the strictest definition of the word, as will be discussed later. Again, the model validation "process" should not be viewed as a mechanism for proving that the model is valid, but rather as a mechanism for enhancing the model, reducing its uncertainty, and improving its predictions through an iterative, long-term, confidence-building process. The process should also contain trigger mechanisms that will drive the model back to the characterization-conceptualization-calibration-prediction loop (i.e., back to square one), but with a better understanding of the modeled system.

Implementing a validation process can help move the modeling project forward beyond the endless loop of characterization, conceptualization, calibration, and prediction, yet will also provide a way back to this loop. However, different parties understand validation in different ways, and there is an urgent need to unify the concepts of model validation and develop a systematic way of testing and evaluating model predictions. This may facilitate acquiring the acceptance of the regulators and the public of the model-based decisions, especially with many sites (e.g., U.S. Department of Energy [DOE] and U.S. Department of Defense [DoD] sites) now having closure processes "knocking on the door" of validation. The development and use of rigorous science to define a process that site sponsors, regulators, and the public can accept will benefit all involved parties.

An actual case that is currently facing the issue of model validation is the Central Nevada Test Area (CNTA), where the Faultless underground nuclear test is undergoing environmental restoration. Underground nuclear test sites are extreme examples of the need for groundwater modeling and for model validation, as a significant radionuclide source will be left in contact with groundwater due to the absence of technically feasible remediation technology. Instead, regulatory closure will depend on a model-generated contaminant boundary (boundary of the area having contaminant concentration exceeding certain threshold) for exercising stewardship restrictions. Confidence in the model results is absolutely critical to achieve closure. A complex, three-dimensional stochastic flow and transport model was developed for the CNTA site (Pohlmann *et al.*, 1999, 2000) and carefully reviewed by the state regulator. Though several aspects of uncertainty were included in that model, concerns remained regarding uncertainty in individual parameter values and the additive effects of multiple sources of uncertainty. A Data Decision Analysis (DDA) was performed (Pohll and Mihevc 2000) to quantify uncertainty in the existing model and determine the most cost-beneficial activities for reducing uncertainty, if reduction was needed. The DDA indicated that though there was large uncertainty present in some model parameters, the overall uncertainty in the calculated contaminant boundary (areas having contamination exceeding a certain standard) during the 1,000-year regulatory timeframe was relatively small. As a result, only limited uncertainty reduction could be expected from expensive characterization activities. With these results, the model sponsor (DOE) and the regulator (Nevada Division of Environmental Protection) determined that the site model was

suitable for moving forward in the corrective action process. Key to this acceptance was the acknowledgment that the model requires independent validation data and that the site requires long-term monitoring (Chapman *et al.*, 2002). Thus, the CNTA model is in immediate need of a validation approach that can stand up to the rigors of scientific peer review, regulatory oversight, and citizen concerns.

Other sites share the need for an effective validation strategy (e.g., the Shoal underground test area, Nevada; Hanford Site, Washington; Maxey Flats, Kentucky; Fernald, Ohio; Oak Ridge National Laboratory, Tennessee; Weldon Springs, Missouri; Nevada Test Site), so that model validation is one of the most critical and challenging issues facing modelers, scientists and regulatory and government agencies. Unfortunately, there does not exist a set of specific procedures and tests that can be easily adapted and applied to determine the validity of a deterministic model, particularly a site-specific model. The validation issue is even more challenging for the "predictive" stochastic models that incorporate effects of parameter uncertainty and spatial variability. As pointed out by Konikow (1986), if a model is to be used for prediction, it should be periodically postaudited, or recalibrated, to incorporate new data and information that may provide different understanding of the processes studied at a certain site. The step of moving forward in the face of uncertainty and proceeding to the validation and long-term monitoring of the Central Nevada Test Area (CNTA) model is consistent with this paradigm since the validation phase and monitoring phase will serve as the periodic postaudit of the CNTA model.

The purpose of this validation plan is to outline a strategy for the different activities that are needed for testing the predictions of the CNTA model. These activities include the field activities for collecting the testing data, the scientific approach that will be used to test the model predictions using these data, the iterative scheme of refining the model and collecting data, and the long-term vision for monitoring the site. Through the validation stage, the focus will be on three major issues: 1) to test how the predictions of the numerical groundwater flow and transport model at CNTA and the underlying conceptual model and assumptions are robust (see definition later) and consistent with the regulatory purposes; 2) to re-evaluate and refine model predictions and reduce the uncertainty level based on data collected in the proposed field activities for this validation; and 3) start the long-term monitoring phase of the site that benefits from and builds on the validation-phase field activities.

In this report, we propose a validation approach for the CNTA model, which addresses some of the important issues that were recognized in previous validation studies, conferences, and symposia as crucial to the process. The proposed approach is an integrated approach that uses a number of tools and approaches for evaluating the predictive CNTA model, refining its predictions, reducing the associated uncertainty, and building the confidence necessary for site closure. The proposed validation methodology focuses on use of collected validation data to reduce model uncertainty and narrow the range of possible outcomes of stochastic numerical models. This requires iterative implementation of data collection, model evaluation, model refinement, and uncertainty reduction. This is particularly critical in radionuclide transport models such as the CNTA model since only a few aspects of the transport modeling results can be tested. This is because the predictions of the model extend thousands of years into the future and no data can be used at this time scale. The key strategy will be to focus on evaluating other model elements (e.g., geologic model, model structure, and flow model) using validation data, which will help refine transport predictions and reduce their uncertainty.

It is important to recognize that the validation issues reviewed in this article are different from many popular model studies that relied on particular field experiments and employed the term "model validation," which referred to validating "process" models or mathematical models.. These experiments, primarily designed for studying and modeling subsurface phenomena, include the Cape Cod experiment (e.g., LeBlanc *et al.*, 1991; Hess *et al.*, 1992), the Borden site test (e.g., Mackay *et al.*, 1986; Freyberg, 1986), the Macrodispersion Experiment (MADE) site (e.g., Boggs *et al.*, 1992), the Twin Lake natural gradient tracer experiment (e.g., Moltyaner *et al.*, 1993), and the Grimsel tracer migration experiments (Frick, 1994). These experiments provided well-characterized sites and reasonably large data sets for calibrating and validating certain process and mathematical models. The common theme was to develop different process models for understanding the physics of flow and transport in the subsurface and use these characterized sites for validating the model conceptualizations and mathematical formulation. Another set of studies focused on calibrating and validating different mathematical models using tracer test results in fractured aquifers (e.g., Maloszewski and Zuber, 1992, 1993; Cacas *et al.*, 1990a, b; Raven *et al.*, 1988; Shapiro and Nicholas, 1989). The scope of these validation studies was to determine whether the values of the model-fitted parameters agreed with those known from independent determinations. The term "model validation" was frequently used in these and other studies and it essentially meant validating a certain mathematical model or verifying the existence of certain processes (e.g., matrix diffusion) using well-characterized field experiments. In addition, the field experiments in these studies were available *a priori*; and models were developed, calibrated, and validated afterwards. For the "predictive" model validation issue we seek to address, validation data must be independent of the characterization and calibration data used to construct the model.

As the two words forming "model validation" have been used with too many different meanings, and since some other terms are interchangeably used for the term "validation," it is necessary to define different terms and to illustrate the intended meaning of these terms when used in this report. The remainder of this report is therefore organized as follows. We present in Section 2 a discussion of the reasons that necessitate the need for validation and the challenges associated with validating a site-specific model such as the CNTA model. In Section 3, we then review the different aspects and definitions of terms such as model, calibration, verification and validation. The purpose of this section is to present clear definitions of and differentiations between the different terminologies as adapted in this report. Model calibration, verification, and validation are thoroughly discussed in this section with a detailed presentation of the discrepancies and the debate in the literature about the meaning and purposes of model validation. Section 4 presents a literature review of the studies and international projects that dealt with model validation issues. This section also discusses the different strategies that were proposed for validating subsurface models (mainly, models of performance assessment of high-level nuclear waste repositories). Section 5 discusses the critical issues and considerations that should be accounted for in developing a model validation plan. Finally, Section 6 outlines a proposed validation plan for the CNTA model with detailed descriptions of some of the underlying theories and hypotheses presented in the Appendices.

2. NEED FOR AND CHALLENGES FACING MODEL VALIDATION

2.1 Need for Validation

Predicting groundwater flow and transport at the field scale is usually done for a specific purpose. A regulatory question arises at a site, for example, and modeling is undertaken to

answer that question. The need for validation arises when the regulatory agency and subsequently the public require assurance that the model's answer to the posed question is a close representation of reality (or at least a conservative estimate). Developers and users of models (i.e., the decision-makers using information derived from model results) and people affected by decisions based on such models are all rightly concerned with whether the model and its results are "correct" (Sargent, 1990). However, depending on the model and the application in question, the correctness requirement may become one of reasonableness. For groundwater models, for instance, all involved parties should be able to understand that the correctness requirement cannot be achieved. Instead, the requirement should shift to good modeling protocol followed by long-term validation and monitoring processes to make sure that the predicted consequences are not underestimated.

As described by Shah Alam (1998), regulators want to be certain that human health and the environment are being protected and general public participation is a key element in a regulatory decision-making process for a contaminated site. Affected public should be able to comprehend and concur with the model on their terms. This is difficult to achieve without a long-term commitment of evaluating and re-evaluating the model results (thus going through a validation process) based on data collected for the validation process and for the long-term monitoring of the site. Most regulators understand modeling well enough to know that a model cannot be proven to be "correct." Rather, they are seeking evidence that the model is sufficient for decision making and that model predictions are being thoroughly tested against site-specific data.

As described by the National Research Council (NRC, 2000), monitoring and validation are needed to improve the understanding of the contaminant fate and transport processes and can be used to recalibrate and revise conceptual and predictive models. NRC (2000) indicates that the ability to monitor and validate is essential to the application of any corrective action to a subsurface contamination problem, but the knowledge and technology bases to support these activities are not fully developed. NRC (2000) thus identifies a number of research needs related to the model validation issue that include the development of validation processes, the development of tools to help judge model performance, and the development of ways to determine the key measurements that are required for the model validation process.

The interest in validating model predictions also arises from the scientific need to better understand the physics of flow and transport in highly complex systems such as the geologic environment. In fact, the invalidated models provide a scientific challenge to researchers to identify the sources of errors in the model and whether these are related to processes that are unresolved or unaccounted for, model structure and conceptual model or input data. In the search for these error sources, new scientific understanding can be gained and progress is usually made by these discoveries.

Aside from these scientific and regulatory motives, and from a relatively legal perspective, the need for subsurface model validation in the U.S. arises from at least two sources (Davis and Goodrich, 1990; Davis *et al.*, 1991). The Code of Federal Regulations states explicitly in 10 CFR Part 60.21(c)(1)(ii)(F) that "Analyses and models that will be used to predict future conditions and changes in the geologic setting shall be supported by using an appropriate combination of such methods as field tests, in-situ tests, laboratory tests which are representative of field conditions, monitoring data, and natural analogue studies." Although this does not call for the strictest application of the term "validation," it does require field tests and

in-situ tests as means for supporting the model (i.e., supporting its use for decision making, which is consistent with our model validation definition presented earlier). The second source quoted by Davis *et al.* (1991) is the legal precedent establishing the need for validation, which was set based on the court case involving the State of Ohio and the U.S. Environmental Protection Agency (EPA) [23 ERC 2091, Sixth Circuit, 1986]. In that case, the court decided that EPA had failed to establish the accuracy of a model that was used for predicting sulfur dioxide emissions from two electric utility plants, as compared with the actual discharge from the plants. The adequacy of the model for its intended use (establishing limitations on sulfur dioxide emissions at the specific power plants) was not checked using the site-specific validation tests.

2.2 Challenges

A number of issues combine to make the validation of subsurface flow and transport models a very difficult and challenging task. First, data are usually lacking for building, calibrating and running the model. Such data are lacking for both simple deterministic models and highly complex stochastic models. Even if there are extensive databases for a particular site, they are often limited with respect to the variety of conditions and parameters that need to be monitored and characterized. With the current level of data scarcity and uncertainty, model validation becomes a formidable task. The question of validation is even more challenging when modeling radionuclide transport thousands of years into the future since no data are available to use for comparison against model predictions at this time scale. In addition, lack of knowledge about future stresses that will affect the groundwater system reduces the reliability of future predictions. Despite these challenges, we need to build confidence that model-based decisions will not result in unacceptable risks to present or future populations or in degradation of the natural environment (Konikow and Bredehoeft, 1992). Building confidence in the models used to support closure of sites is the requirement for validation; developing a validation process that allows regulatory closure of sites with significant groundwater contamination should, therefore, be the ultimate goal of any validation strategy.

Konikow and Bredehoeft (1992) further argue that the only solution to the above challenge is the notion that our fundamental understanding of the processes encountered in the subsurface will help make defensible long-term predictions. Expert judgment and the approval of the scientific community come into the picture under these challenges. Models, however, serve to sharpen our professional judgment and increase our understanding of the very complex subsurface systems. Heterogeneity is another challenge when it comes to prediction and validation. Heterogeneity makes it difficult to fully characterize the subsurface, especially with the difficulty of making subsurface observations. When heterogeneity is significant and data are limited, as is the case in many field sites, there may be no objective way of judging the model predictions or declaring any degree of satisfaction about the model. It is also important to note that even if we can get a highly detailed and reasonably accurate characterization of the subsurface parameters, the validation process may still be very difficult. If one tries to obtain a detailed prediction of some heterogeneous variables such as the groundwater velocity or the contaminant concentration, it may be impossible to collect enough data to verify whether or not the predictions are correct. On the other hand, we may be satisfied with reliable descriptions of larger-scale trends such as averaged velocity or concentration over a specified volume and/or time interval. Unfortunately, it may be difficult to estimate such trends from the limited numbers of point measurements, which are typically collected in field experiments (McLaughlin and Luis, 1990).

A major difficulty has also been deciding on quantitative criteria on which to base the decision that there is "agreement" between predicted and measured values (Voss, 1990). Furthermore, uncertainties inherent in describing and modeling complex natural systems make it difficult to discriminate between inadequacies in the conceptual models, mathematical models, and input data.

Another challenge for model validation is the high cost of obtaining data for testing the, which should be considered in designing any validation plan. There is a limit beyond which increased investment in model validation efforts (both data collection and analysis) does not significantly increase confidence in the model and adds little value to the end user (Sargent, 1990). Therefore, the model validation process requires consent between concerned parties regarding the level of confidence required for the model to be validated, keeping an eye on the cost that is needed to achieve this confidence level. This type of consent or agreement may be difficult to attain, as there may be conflict of interests or disagreement on the meaning, objectives and purposes of model validation among the parties involved in the process.

3. REVIEW OF TERMINOLOGY AND DEFINITIONS

A first step in developing the validation methodology is to define the meaning intended for different terms related to model validation and to review previous efforts and strategies of model validation. Validation, verification, and confirmation are all concepts in terms of groundwater numerical models that not only do not have established and generally accepted practices, there is not even widespread agreement on the meaning of the terms as applied to models. It is, therefore, important to define the different terms used in the literature and to illustrate the intended meaning of these terms when used in this report. In addition to these definitions, we present in this section the different arguments and the debate about the meaning of the term "validation" as it applies to groundwater models.

3.1 Models

A model is simply an abstraction or a simple representation of a real system or process. One can distinguish between three types of models: conceptual, mathematical, and numerical. A conceptual model can be defined as a hypothesis for how a system or a process operates and is qualitative in nature. This operation can then be expressed quantitatively as a mathematical model. Mathematical models are abstractions that replace objects, forces, and events by expressions that contain mathematical variables, parameters, and constants (Konikow and Bredehoeft, 1992). When the mathematical model is implemented via a computer code to perform the actual model computations, the numerical model for the problem at hand is established.

For predicting groundwater flow and transport at field sites, models have to include a number of components: (1) a conceptual model of flow based on geologic, hydrologic, and chemical information, (2) a mathematical flow model expressing the processes affecting the flow system (e.g., recharge, source/sink terms), (3) a computer code incorporating the mathematical model of the flow system, (4) a conceptual transport model based on the definition of the contaminant source, release scenarios, and transport properties of the subsurface environment, (5) a mathematical transport model, and (6) a computer transport code for solving the mathematical transport equations.

One can also distinguish between "generic models" and "site-specific models." The computer codes that are used to solve the mathematical flow and transport equations are referred to as generic models, whereas after combining them with the conceptual models (model structure), input data and boundary conditions for a particular geographical area, they become site-specific models. These site-specific models thus rely on four components to predict flow and transport: model structure (conceptual models), initial and boundary conditions, the input data, and the computer code. The inadequacy of any of these items or their nonconformity with the real system will most likely be an obstacle to accepting the model as the basis for making decisions. However, as will be seen later, the adequacy of these items does not necessarily mean a "valid" prediction.

In terms of their use, models can be classified into two types: research or analysis models and predictive or decision-making models. Research models are common in studying and understanding different phenomena in the subsurface and they usually rely on hypothetical domains or well-characterized field sites. Many of the field experiments that were covered in a number of international workshops and symposia on model validation (see next section) can be described as research or analysis models. These models were focused on understanding a number of transport issues, e.g., matrix diffusion in fractured media and kinetics of sorption in the fractures and the surrounding porous blocks. Predictive models, on the other hand, are mainly used to support and aid a regulatory decision regarding a subsurface contamination issue. Performance assessment models of high-level nuclear waste repositories, predictive models of radionuclide transport associated with these repositories and models of nuclear testing sites belong to the category of predictive models.

3.2 Model Calibration

In the earth sciences, the modeler is commonly faced with the inverse problem: the distribution of the dependent variable (e.g., head) is the most well-known aspect of the system, whereas the distribution of the independent variable (e.g., conductivity, porosity) is the least well known (Oreskes *et al.*, 1994). Model calibration is the process used to solve this inverse problem. That is, model calibration is the process of tuning the model to identify the independent input parameters by fitting the model results to some field or experimental data, which usually represent the dependent system parameters. The calibration process can be quantitatively described by the goodness of fit. When the model is used for long-term prediction (e.g., 1000s of years at underground nuclear testing sites), it is often calibrated using short-term data. This calibration cannot replace validation, but can only be considered as part of the site characterization and model formulation process. In some situations, the validation task may become one of a calibration, whereby experimental data that are collected for the validation purpose are used during the modeling effort. Although this type of calibration builds some confidence in the model results, especially if the calibration fit is good, calibration by itself is not validation because the input parameters of the model are found based on the experimental results that can no longer be considered as validation data (Davis *et al.*, 1991). Furthermore, a good calibration of a model does not necessarily imply that the model is valid beyond the values and conditions of this calibration. Therefore, field or laboratory experiments for model validation studies should follow the model simulations to ensure that the validation effort is not simply a calibration effort based on those experimental results (Davis *et al.*, 1991).

Anderson and Woessner (1992a) point out the need to distinguish between calibration, verification and validation, while realizing that the three processes are simply tests of model

accuracy. In their description, calibration is a trial-and-error adjustment of parameters that can be done manually or using an automated parameter estimation code, whereas model verification is aimed at establishing a greater confidence in the model by using a set of calibrated parameter values and stresses to reproduce a second set of field data. Verification can be used as part of the modeling protocol for testing the governing equations, the numerical model, and the code. Konikow (1986) also defines a verified model as the model for which the accuracy and predictive capability have been proven to lie within acceptable tolerance using tests independent of the calibration data. According to Anderson and Woessner (1992a), model calibration and verification demonstrate that the model can mimic past behavior, whereas model validation tests whether the model can predict the future, which they call a predictive validation or postaudit. They further assert that this type of postaudit should be performed a long time after the initial calibration and prediction are made so as to allow the system represented by the model to evolve away from the calibrated state. With a more or less similar view, de Marsily (1990) states that the calibration and validation of groundwater flow models at sites characterized by low permeability media must take into account that the present observed state of the groundwater system may be the result of a long transient history.

3.3 Code Verification

There should be a clear distinction between code validation/verification and model validation. A computer code is said to be certified if the code is properly verified and properly documented (Tsang, 1991). Verification of a mathematical model or its computer code is obtained when it is shown that the model behaves as intended, i.e., that it is a proper mathematical representation of the conceptual model and that the equations are correctly encoded and solved (Maloszewski and Zuber, 1992). Tsang (1991) argues that it is illogical to use the term "code validation," as "validation" questions the appropriateness of the mathematical equations and input data and conditions, which are assumed and taken for granted in a code. A code can only be certified or tested, but not validated. Validation becomes an issue for a model that is developed to answer a site-specific question.

Taking a more philosophical view, Oreskes *et al.* (1994) define a verified model as one whose truth has been demonstrated and argue that it is impossible to demonstrate the truth of any proposition except in closed systems. They distinguish between mathematical models that may be verifiable, just as an algorithm within a computer code may be verifiable, and between the models that use these mathematical components, which are never closed. These models contain unknown elements that modelers conceptualize based on expert judgment and require input parameters that are incompletely known. They postulate that verification is only possible in closed systems in which all the components of the system are established independently and are known to be correct. They go on to demonstrate that if model results compare unfavorably with observations, then it can be concluded that something is wrong in the model, but if the comparison is favorable, a dilemma exists in judging the model. If two or more errors cancel each other out, there is no way to know that this cancellation has occurred and a faulty model may be seen as correct. Their bottom line is that a good match between predicted and observed output does not verify an open system.

Van der Heijde and Kanzer (1997) address in a great detail the issue of code testing for groundwater problems. They focus on testing the code functionality and performance using benchmarking with known, independently derived solutions, intracomparison using different code functions inciting the same system responses, intercomparison with comparable simulation

codes, and comparison with field and laboratory experiments. Along similar lines, Beljin (1988) uses three levels of model testing for evaluating solute transport models in two dimensions. The first level uses analytical solutions to verify the numerical technique and illustrate the behavior of the numerical solution. The second level includes hypothetical problems and examines such aspects as the model's response to aquifer heterogeneity, anisotropy, and irregular boundaries. The last level involves history matching with field data.

Konikow and Bredehoeft (1992) also distinguish between code verification and model validation. They postulate that the former essentially answers the question: Does the computer code provide an accurate solution to the governing partial differential equation for various boundary value problems? This can be demonstrated by showing that the code gives good results for problems having known solutions, which are very simplified problems. However, after adding the different complexities to the code in addressing a certain site-specific problem, the question becomes how to prove that the code still gives an accurate solution to the governing equations under these complex conditions, for which no analytical solution is available. Konikow and Bredehoeft's (1992) answer to this question is that there is no way of assuring such accuracy, but only checking simple aspects such as mass balance. We agree that this strict validation perspective is unachievable given our current knowledge and technology levels.

Another set of studies focus on using laboratory experiments to verify certain mathematical equations or constitutive relationships. As an example, Hassanizadeh (1990a) used a set of laboratory column experiments to investigate some of the relevant processes in brine transport in porous media and to provide partial data sets for validating (actually verifying) different forms of Darcy's law and Fick's law for density-driven conditions. In his study, the experimental (or validation) data were available before selecting the appropriate form of Darcy's law and Fick's law to better describe the experimental data. This is in essence completely different than the validation process for site-specific, predictive models.

The ASTM guide (ASTM, 1993) also distinguished between application verification (or site-specific model validation/evaluation) and code verification. The former refers to the process whereby a model, its computer code, boundary and initial conditions are tested by simulating independent data from different hydrologic conditions to establish the predictive capability of the model (Johnson and Weimer, 1996), whereas the latter refers to software testing, comparison with analytical solutions, and comparison with other similar codes to demonstrate that the code represents its underlying mathematical foundation (ASTM, 1993).

So in summary, the term "verification" should refer to the demonstration of the ability of a generic model (and maybe an analysis model) to solve the governing equations, whereas validation should represent the process of post-prediction testing and evaluation of a site-specific model for the purpose of supporting the decision making that relies on modeling results. When data are available to split between calibration and "verification," it is common to call the process of using the calibrated model to reproduce the "verification" data set a model verification process. This process is different from the model validation process as it is part of the development stage of the model, and apparently, the modelers can and do change the model conceptualization if the calibrated model fails to reproduce the verification data set. Model validation process comes after the completion of this loop and is aimed at building confidence in model predictions that are going to be the basis for decision making.

3.4 Model Validation

The term validation is featured prominently in the literature on high-level radioactive waste disposal. Pescatore (1994) reports that there is a lack of use of the term validation in the field of low-level radioactive waste disposal and also, during the first half of the last century, in all technical fields. The first technical appearance dates from the mid-1950s and it was adopted thereafter in the computer field and elevated to its present status following the computer revolution in the 1970s and early 1980s (Pescatore, 1994). The term validation then started to appear in some high-level waste safety standards in the late 1980s. A large number of definitions exist for the term "validation" within the performance assessment community, and it has been used with many different meanings, sometimes in the same report.

Most of the controversy over validation arises from alternative interpretations and perceptions of the meaning of the term. Interpretations range from an inherently unachievable "proof of truth" to more pragmatic approaches in waste management with emphasis on the subjective assessment of whether models are "good enough" for the application at hand (Zuidema, 1994). Different types of classification or categories for the numerous definitions of model validation have been presented in the literature. Here, we combine these classifications and categorize the validation definitions and perspectives into four categories. The following four subsections summarize the different definitions and the controversy in the literature regarding the meaning and objectives of model validation.

3.4.1 Scientific Views of Model Validation

The dictionary definition of *valid* covers a wide range of meanings (e.g., strong, having sufficient strength or force, sound, effective, convincing, fulfilling all necessary conditions, founded in truth, logically correct, executed with the proper formalities, having such force as to compel serious attention). The scientific view of validation usually implies that models are "true" representations of reality. The U.S. Nuclear Regulatory Commission (USNRC, 1984) defines validation as the process of obtaining assurance that a model, as embodied in a computer code, is a correct representation of the process or system for which it is intended. DOE (DOE, 1986) defines validation as a process to ascertain that the code or model reflects the behavior of the real world. Niederer (1990) argues that three validation types can be invoked: (1) the Popperian (after Popper, 1968) approach of falsifying wrong theories (more of a philosophical view), (2) the positive proof approach, which is partly achieved by showing that the theory (or model) is able to explain pertinent observations and experimental data, as this ability is a necessary but not sufficient condition, and (3) the consensus-based approach proposed by Kuhn (1970), who concludes that proof of a scientific theory largely rests on consensus. For a scientific theory, consensus-based validation means that acceptance is ultimately based on the feeling that the theory works, a feeling that grows from repeated successful use. In terms of groundwater model validation, the latter type calls for providing ample positive evidence for the appropriateness of the model, which will lead to general consensus that the model is adequate. Niederer (1990) states "consensus is one aspect of scientific truth… However, as far as public acceptance is concerned, it is the only one that really counts." Jackson *et al.* (1990, 1992) argue that validation should be different for general models and specific models. That is, for a general model, validation consists of establishing the case for the model such that the model is widely accepted within the scientific community. But for specific models, validation consists of establishing a case such that one might reasonably expect someone with relevant technical knowledge would consider the model acceptable. Jackson *et al* (1990, 1992) consider validation to be about

establishing whether or not the model is an acceptable representation of the physical system and checking that the model is internally consistent and consistent with principles that are generally accepted in the scientific community.

Anderson and Woessner's (1992a) approach to validation has a slightly different perspective. They define the strictest form of validation as the demonstration of the model's ability to accurately predict the future, which they call a postaudit. Later, Woessner and Anderson (1996) provided a less stringent requirement for accepting groundwater models and indicated that this acceptance should be based on confirming observations to support a subjective judgment. They also emphasized the importance of understanding the role of uncertainty and accepting it when dealing with groundwater modeling.

The central problem with the language of validation and the strict definition of model validation, as seen by Oreskes *et al.* (1994), is that it implies an either/or situation, but in practice only a few (if any) are entirely confirmed by observational data, and a few are entirely refuted. In addition, both terms are affirmative and they encourage the modeler to always claim a positive result, which is the reason it is impossible to see a sentence like "the observed data invalidates our model" in published modeling studies (Oreskes *et al.*, 1994).

The scientific views of model validation are more suitable, if necessary at all, for theories and mathematical developments that need to be validated in a strict sense. For numerical groundwater models that are used to support or guide a decision-making process related to a subsurface problem, these scientific views are essentially neither achievable nor relevant. Accuracy is not always required for using model results as a basis for decision making. If, for example, one monitors a noncontaminated area as delineated by a model, one would only try to make sure that the clean area is in fact clean regardless of whether the model accurately predicts how contaminated the area within the plume is (concentration values).

3.4.2. Philosophical Views of Model Validation

According to Konikow and Bredehoeft (1992), philosophical definitions of validation are based on two different views. The first of these argues that theories are confirmed or refuted based on the results of critical experiments designed to verify the theory consequences. The second philosophical perspective is that as scientists, we can never validate a theory or a hypothesis but can only invalidate it. Popper (1968) states that a model, theory, or hypothesis can never be proven to be true, no matter how much corroborative data are presented; they can only be falsified. Consistent with the first view, Schlesinger (1979) defines validation as meaning "substantiation that a computerized model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model." However, Konikow and Bredehoeft (1992) believe that many, if not most, present-day scientists who have considered these issues find themselves in agreement with the second view. They also add that groundwater models are subject to improvements via invalidation, but cannot be proven valid and that validation cannot add to the fund of knowledge.

The philosophical view of model validation articulated by Oreskes *et al.* (1994) is that the term does not denote establishment of truth but rather legitimacy. They define a valid model as one that does not contain known or detectable flaws and is internally consistent. They, however, agree with the common view that the establishment of a model's ability to "accurately" represent the actual processes occurring in the real system is not even a theoretical possibility. Extending their views, Oreskes *et al.* (1994) use the term *confirmation* to account for the fact that a failure

to reproduce observed data falsifies the model, but the reverse is never the case. By using as numerous and diverse confirming observations as possible, it is likely reasonable to conclude that the conceptualization embodied in the model is not flawed. It is important, however, to recognize that confirming observations do not demonstrate the veracity of a model or a hypothesis; they only support its probability (Oreskes *et al.*, 1994).

Bredehoeft and Konikow (1993) argue that using the terms "validation" and "verification" are misleading as they imply the correctness of the groundwater models, which none of the groundwater modelers would claim. They suggest that the groundwater community should abandon these terms, and that the term "history matching" used in petroleum engineering be used instead. This term encompasses the processes of calibration and validation without connotation of correctness. They, however, caution that care should be taken to predict only for a time comparable to the period that was matched. McCombie and McKinley (1993) argue against these views and assert that the key problem in the validation issue is to define what level of accuracy and what degree of confidence must be achieved in the prediction of specific parameters. The decision about how much effort must go into the validation process before the model can be considered to be valid is necessarily subjective and very dependent on the complexity of the system and on the objective of using the model in the first place. McCombie and McKinley (1993) further recommend that the subjective aspect of assessing if a model is good enough be included in the term "validation." de Marsily *et al.* (1992) present evidence from their modeling experience as proof that their groundwater model has been validated or at least proven to be not invalid.

In an attempt to reconcile these two opposing view points, Leijnse and Hassanizadeh (1994) postulate that Konikow and Bredehoeft (1992) and de Marsily *et al.* (1992) refer to two slightly different, yet related, definitions of the terms "model" and "validation." They state that de Marsily *et al.* (1992) invoke a weak definition of the word "model," wherein the mathematical equations and simplifying assumptions are included but not the input data. On the other hand, Konikow and Bredehoeft (1992) invoke the strong definition of the word "model" where all the above components are included in addition to the parameter values, boundary conditions, system geometry, and sources and sinks. Parallel to these definitions, the validation may be viewed in a weak sense and in a strong sense. Validation in the weak sense refers to the validity of the conceptual part of the model (Leijnse and Hassanizadeh, 1994), and is applicable to models that are used in an analysis mode to analyze a system of interest and to increase understanding of its behavior. The strong definition of validation as discussed by these authors implies the validity of the model of a given system as a whole, including all input data, which is related to the predictive ability of a model in mimicking the right system behavior. When Konikow and Bredehoeft (1992) say that 'groundwater models cannot be validated,' they refer to validation in the strong sense and they have 'prediction models' in mind (Leijnse and Hassanizadeh, 1994).

In Konikow and Bredehoeft's (1992, 1993) terms, the finding that the model is not proven to be invalid does not mean that it is valid. This is true for validation in the strictest sense, but this author believes that for practical purposes, decision-making purposes, and for moving forward toward better understanding, the model success at the invalidation attempts means that the model is successfully progressing through the process of model validation. The terms suggested by Konikow and Bredehoeft (1993) are mostly helpful in the model development, building, testing, and usage stages. If the model is accepted by a regulatory agency, then the process has moved beyond these terms and the more relevant term is in fact model validation. When it comes to the

public perception of the term "validation," the term requires as much effort in explanation as do terms like "calibration," "history matching," and "benchmarking." For example, to the layman, the term "calibration" gives the same indication of accuracy and correctness as the term "validation." Thus, using and explaining the real meaning of the model validation process should not be any different than using and explaining any alternative term.

Most models, if not all, are not being used to reveal the truth of a system. Of course it would be great if models could do so, but they simply cannot. Models are in many cases decision-making tools. When a model successfully passes a rigorous development, calibration, and testing process, it becomes a reasonable decision-making tool given the limited data used in the development process. Acknowledging the role of uncertainty, the model-validation process is one crucial stage in the entire process that should be regarded as an additional filter for independent model evaluation. The fact is that most of the literature debate is on the terminology and not on the process itself. No one argues that the process is unimportant, unneeded, or useless and no one disagrees on the concept of using an independent data set to test the model. The disagreement is in what we call it and what the implications are for the term we use.

Other philosophical views were presented more recently in a series of articles edited by Anderson and Bates (2001) focusing on model validation perspectives in hydrological sciences. For example, Young (2001) states that the views articulated by Konikow and Bredehoeft (1992) and Oreskes *et al.* (1994) are linked in part to questions of semantics: what is the truth? What is meant by terms such as validation, verification, and confirmation? etc. Young (2001) also postulates that when models are not proven to be false, they can be considered conditionally valid in the sense that it can be assumed to represent the best theory of behavior currently available that has not yet been falsified.

In an interesting editorial following the above-cited articles, Bair (1994) presents a personal experience from the courtroom, where the groundwater model validation issue was the key element of a $500 million lawsuit. He was testifying in that trial and was asked to evaluate the plaintiffs' and defendant's groundwater models, which predicted migration of contaminants for 17,000 feet and 5,000 feet from the injection point, respectively. During the different phases of the trial, the plaintiff's attorney used Bredehoeft and Konikow's (1993) arguments that groundwater models cannot be validated, only invalidated. Bair (1994) mentioned that his response, which was against this argument, was supported by McCombie and McKinley (1993) and de Marsily *et al.*'s (1992) comments on Bredehoeft and Konikow's (1993) arguments. Summarizing Bair's (1994) conclusions about this experience, it shows how the jury understood the difference between predictions that are certain beyond reasonable doubt (operational or confidence-based validity) and predictions that are certain beyond any doubt (strictest form of validity). The doubts were probably removed from the jury's mind by the amount of site-specific data, the small differences between measured and simulated pressures and concentrations, and the recognition that no data were presented that invalidated the defendant's model. The reasonableness of a modeling effort can only be supported by a large number of confirming observations that remove reasonable doubt (Woessner and Anderson, 1996).

The main point one can capture from the above discussion is that we should use as much data as possible to try to invalidate a model to show that it is either certain or uncertain beyond reasonable doubt. Reasonable doubt is fundamental to the scientific method, but should not prevent us from making predictions; it should cause us to gather sufficient data to rigorously test our models so that we make well-founded predictions (Bair, 1994). For performance assessment

models of nuclear repositories, the regulators in both the United States and Sweden require "reasonable assurance" that the models comply with regulatory criteria. This concept recognizes that absolute assurance of compliance is neither possible nor required, but model developers should provide such information as may be necessary to convince a "reasonable decision-maker" that compliance with regulatory criteria would be achieved (Eisenberg *et al.*, 1994). However, if such an approach is assumed for validation, there can be no standard answer to the question "How much validation is enough?"

3.4.3 Operational Views of Model Validation

A number of operational definitions consider validation from a practical and regulatory perspective. From the practical perspective and in the context of groundwater flow models, the ASTM (1993) and Brown and Laase (1995) define model validation (or application verification) as the process of using a calibrated model to approximate acceptably a second set of field data measured under similar hydrologic conditions. Van der Heijde (1990) states that the objective of model validation is to determine how well a model's theoretical foundation and computer implementation describe the actual system behavior in terms of the degree of correlation between model calculations and independently derived observations of the cause-and-effect responses of the actual groundwater system.

The International Atomic Energy Agency (IAEA, 1982) defines validation to be attained when a conceptual model and its associated computer code provide a good representation of the actual processes occurring in the real system. However, depending on the meaning and strength of the term "good representation," this definition may become a scientific one as opposed to a practical/operational one. Flavelle (1992) argues that most of the validation definitions available in the literature make explicit reference to the need to demonstrate that a model is a good, correct, or sufficient representation of reality and that these definitions require subjective interpretation but do not recognize the need to measure the accuracy of the model calculations. He, therefore, adds establishing the accuracy of the model predictions as a second dimension to the definition of validation. In his argument, Flavelle relies on an updated validation definition provided by the IAEA (IAEA, 1988), which describes the requirements for validating a model as "a model cannot be considered validated until sufficient testing has been performed to ensure an acceptable level of predictive accuracy. (Note that the acceptable level of accuracy is judgmental and will vary depending on the specific problem or question to be addressed by the model)." Borgorinski *et al.* (1988) follow the IAEA's definition that model validation is confirmed when the model provides a good representation of the actual processes that occur in reality.

It is clear that these operational definitions rely on a subjective component in the judgment of a model's validity. However, any evaluation process for model aspects, including the calibration evaluation, has to rely on subjective judgment. There is no unique way to structure a process for attaining a reasonable evaluation or guiding the subjective element of the validation process. Since subjectivity will always be complemented by objective, quantitative analysis, the balance between the two aspects depends largely on the problem at hand and the risk associated with making an unacceptable (or bad) judgment. With the significant progress made during the past three decades in regards to the understanding and acceptance of the role uncertainty plays in groundwater models, it should not be considered a weakness that subjective judgment and hydrogeologic expertise are integral components of the entire modeling (including validation) process.

Tsang (1991) defines validation as follows: "a model, including the conceptualization and the code, can be said to be validated with respect to (a) a process or (b) a site-specific system." He argues that one should carry out model validation with respect to various processes, and at a certain site, one should identify the relevant processes and the model geometric structure and carry out the validation of the model (or group of models) with respect to that specific site. Tsang (1991) further states "it is illogical to refer to a validated model in the generic sense, but it can be stated that a model is validated with respect to a given process, or a group of models are validated with respect to a given site." Voss (1990) describes model validity as the process of showing that the model is appropriate and adequate for the problem being addressed, is logically developed using the best available technology, is supported by high quality experimental and observational data, and that the limitations of the model are well understood.

For performance assessment of nuclear repositories, McCombie *et al.* (1990) discuss the interplay between achieving robust models and validated models. They argue that a model is determined to be robust when there is confidence that any errors will either have little effect on performance or be on the conservative side. However, they also emphasize that we should always aim at achieving the best possible understanding of system behavior and a realistic modeling of all the important processes involved. Along similar lines, Zuidema (1994) and Frick (1994) support the idea that it is not critical that models are strictly correct and include all natural details and processes, but that any uncertainty and simplification results in overestimating the consequences (conservatism). It is important to recognize that the conservative assumptions employed in the modeling process because of lack of data can, at a later stage when more information is available, be replaced by more realistic representations. The ignored phenomena are thus marked as "reserve phenomena" and may be included in the model at a later stage (Zuidema, 1994).

The concept of conservatism and the requirement that models need to go through a validation process represent two processes aimed at assuring the public that decisions made based on model results will not compromise the public health and safety. However, both processes may not actually change the preconceived public perception about contaminated groundwater close to residential areas. For example, to a layman living close to a site where drinking water may pose a health risk, it does not matter how much conservatism modelers build into the model. For those living far enough from such a place, again it does not matter that much what value of risk the contaminated groundwater poses for human health. Thus, the public perception is determined *a priori* regardless of whether modelers use many conservative assumptions and whether they use the term "model validation," "benchmarking," "history matching," etc. The point is that using the term "model validation" will not mislead any of those highly concerned about the problem at hand. For those who are amenable to independently evaluating and understanding the modeling process, any used term and the underlying limitations will have to be adequately simplified and explained.

3.4.4 Confidence-building Views of Model Validation

Davis and Goodrich (1990) identify two acceptance criteria for a given model. The first is a measure of the adequacy of the model structure (conceptual model, mathematical model) in describing the system behavior and the second is a measure of the accuracy of the model input parameters relative to experimental results and field observations. Along similar lines, Luis and McLaughlin (1992) postulate that model validation addresses the question of whether or not a model adequately represents observed phenomena (qualification of the model), whereas accuracy

assessment addresses the larger question of how well a model will perform under conditions that have not yet been observed. They view model validation (i.e., a comparison between model predictions and observations) as a first step that establishes the ability of the model to explain observed phenomena. If the model passes a set of reasonable validation criteria that build confidence in its performance, one can then proceed with an accuracy assessment, which assumes that structural errors (stemming from conceptual model, mathematical formulation and computer code) are negligible (Luis and McLaughlin, 1992) or are captured in the overall uncertainty range. Although conceptual errors will always remain unknown, if the problem is cast in a stochastic framework with uncertainty considered in different parameter values, small conceptual errors can be considered minor relative to the entire range of uncertainty. However, for major conceptual errors (e.g., not accounting for matrix diffusion in a fractured system, neglecting vadose zone processes in a saturated-unsaturated system, etc.) the model will most likely not pass any rigorous set of tests and evaluations.

Neuman (1992) defines the validation of safety assessment models as the process of building scientific confidence in the methods used to perform such assessment, and recognizes, however, that this confidence-building approach to validation is possibly open-ended, as many iterations between modelers and regulators as may be needed. Eisenberg *et al.* (1994) support the idea of confidence building and indicate that this term recognizes that full scientific validation of models of performance assessment may be impossible and that the acceptance of mathematical models for regulatory purposes should be based on appropriate testing, which will lead to a reasonable assurance that the results are acceptable. Hassanizadeh (1990b) differentiates between two types of validation efforts. The first is research (or analysis) model validation and the second is safety assessment modeling (or predictive modeling) validation. The research model validation is a tool that helps one understand processes, uncertainties, etc., whereas the safety assessment modeling validation or predictive validation is a goal that helps the decision-making process. Sargent (1990) regards validation as a process that consists of performing tests and evaluations during model development to determine whether a model is valid or not. Several models or versions of a model are usually developed in the modeling process prior to obtaining a satisfactory valid model. Tests and evaluations are conducted until sufficient confidence is obtained that a model can be considered valid for its intended application.

In validating a nitrate percolation model, Mummert (1996) indicates that because of errors in data collection, input parameters, conceptual and parameter uncertainty, it is difficult to see how any groundwater model can be shown to be completely valid in the strictest sense of the term. Validation in Mummert (1996) is thus presented as the process of building confidence in the model rather than determining its absolute correctness or incorrectness.

Both opponents and proponents for the term “validation” agree that the main concern is one of adequacy and not correctness. That is, the main concern is always whether or not the model is adequate for its intended use and whether or not there is sufficient evidence that the model development followed logical and scientific approaches and did not fail to account for important features and processes. Also, it should be noted that determining the adequacy of a model or building confidence in its prediction is not a one-time exercise. It is an iterative process that should be viewed as part of an integral loop with trigger mechanisms or decision points that force the process back to the characterization-conceptualization-building-calibration-prediction loop if the model adequacy tests (or model validation process) indicate the need to do so.

3.5 Model Inadequacy

There are two main reasons for the failure of a model to adequately represent a physical system: 1) the general physical laws underlying the model may be inappropriate for the problem at hand (e.g., using a porous medium assumption for a fractured system where matrix diffusion is a strong phenomenon), or 2) the representation of parameters in the model are inappropriate (e.g., ignoring spatial variability). It is important to re-evaluate the model and try to identify the sources of the mismatch between observed and predicted system behaviors. This is important as it helps isolate the possible sources of error and quantify the contribution of each source to the total error or mismatch. Three general sources of errors can exist: errors in the conceptual model itself, numerical errors arising from the solution of the mathematical equations, and errors arising from the uncertainties in the input parameters. Tsang (1994) adds the possibility that the field data used in validation may not be representative of the real system that is being modeled. No matter how sophisticated the modeling approach is, when applied to a subsurface flow and transport problem, it provides no more than a simplistic representation of very complex field conditions. In this regard, Konikow (1986) states that models should be considered as dynamic representations of nature, subject to further refinements and improvements. As new data become available (e.g., through new wells), model predictions can be evaluated, validated or invalidated, and then modified if necessary.

3.6 Discussion

The above definitions cover a wide scope of different views on groundwater model validation. Nevertheless, many of these definitions focus on providing evidence that the model under consideration is adequate for its intended use. We postulate that model validation is a long-term, iterative process aimed at building confidence in the model as a whole and with trigger mechanisms that drive the process back to the beginning if major deficiencies are found. Key to this process is the use of a diverse set of tests that should be designed to evaluate a diverse set of aspects related to the model.

Overall, there is a general agreement between different definitions of validation, which is centered around the fact that absolute proof that models are perfect representations of reality is usually not required. Adequate representation of reality is what most of the validation definitions focus on. However, a subjective judgment will eventually answer the subsequent question of "how adequate is adequate?" Davis *et al.* (1991) mention that the determination of the model adequacy should consider the types of validation tests, the number of validation tests, the degree of agreement between model and the validation tests and the conformity between model descriptions and site-specific information. They emphasize the necessity for rigorous development of the validation process and the importance of providing regulators with validation information that is as inclusive as possible and follows a logical systematic approach.

An important point here is that the tests of model predictions should be suitable for the regulatory purpose of undertaking the modeling exercise in the first place. That is to say that the validation tests should not be focused on whether the model is scientifically correct for all conditions, but rather on the adequacy of the model for the intended regulatory purpose. For example, a transport model predicting the spatial-temporal distribution of contaminant concentrations can be impossible to validate in terms of matching measured and predicted concentrations, but can be validated or invalidated from a regulatory perspective (e.g., whether or not the plume will reach a certain compliance boundary within a certain time frame). In this

case, the model validation exercise should be tailored to this purpose. That is, if the model predicts that the plume will not reach that boundary, experimental evidence is needed to support that prediction regardless of the mismatch between predicted concentrations and field-measured concentrations. What is important in this case is whether any field evidence indicates faster (or slower) transport rates, thus earlier or later mass arrival, than predicted.

It is clear that often-quoted statements such as "groundwater models cannot be validated" and "groundwater models can only be invalidated" (Konikow and Bredehoeft 1992, 1993; Bredehoeft and Konikow 1992, 1993) refer to validation in only the strictest sense, responding to a concern that the layman's possible misconceptions are of predominant importance. Unfortunately, such statements may lead to a laid-back attitude on the part of researchers, consultants, and even regulatory agencies when it comes to evaluating model predictions. With the perception that no matter what we do, the groundwater model will never be validated, temptations are high that good model development, building, and calibration are the end of the story and nothing can be done more than a monitoring well placed downstream, even though the downstream direction itself may need to be validated. All groundwater modelers agree that their models cannot be validated in the strictest sense (at least with present-day technology), but similarly agree on the importance of post-prediction testing and evaluation. By expanding our definition of validation to encompass a long-term process of confidence building, modelers and model users can develop rigorous validation processes that will ultimately improve model quality and the quality of decisions based on models.

With the state of current knowledge and technology, we believe that the operational and confidence-building definitions of model validation are more amenable to implementation and practicability, especially when predictions are obtained for thousands of years into the future. The definitions essentially lead to an iterative process that is aimed at adjusting model conceptualization, structure, and input as new data become available in such a way that reduces prediction uncertainty and builds confidence in the model results. Furthermore, when there are insufficient data to split between calibration and short-term validation, one has to answer the regulatory concern based on model results and a limited number of field activities that should be carefully designed in light of the understanding of the system behavior provided by the model and available data.

4. REVIEW OF GROUNDWATER MODEL VALIDATION STUDIES

Research on model validation is extensively reported in the literature, but unfortunately does not provide a quantitative approach or outline a step-by-step procedure for achieving any type of model validation. In this section, we review some of the major studies, international conferences, projects, and symposia that were mainly devoted to address the issue of subsurface model validation. In the area of toxic waste management, a number of authors (e.g., Moran and Mezgar, 1982; Huyakorn *et al.*, 1984; van der Heijde *et al.*, 1985; van der Heijde, 1987; Beljin, 1988) have considered the question of whether a model used in a safety assessment program is valid in making appropriate long-term predictions. In addition, during the late 1980s, an effort was made to establish a groundwater research data center for the validation of subsurface flow and transport models (Miller and van der Heijde, 1988; van der Heijde *et al.*, 1989).

In the area of nuclear waste management, the need to validate groundwater models has received increased emphasis. This has led to institutionalized and publicized programs for validation of hydrogeological models. A number of international cooperative projects such as

INTRACOIN (1984, 1986), HYDROCOIN (Grundfelt, 1987; Grundfelt *et al.*, 1990), INTRAVAL (Andersson *et al.*, 1989; Nicholson, 1990), STRIPA (Herbert *et al.*, 1990), CHEMVAL (Broyd *et al.*, 1990), BIOMOVS (SSI, 1990) were devoted to the validation of models. Model validation was also extensively discussed in symposia including GEOVAL87 (1987), GEOVAL90 (1990) and GEOVAL94 (1994). The journal *Advances in Water Resources* dedicated two special issues to the topic of model validation (AWR, 1992a, b). Additionally, a wealth of literature has been published on validation in the field of systems engineering and operations research (Tsang, 1991), some of which may be useful for subsurface model validation. Examples cited by Tsang (1991) include Balci (1988, 1989), Balci and Sargent (1981, 1982, 1984), Gass (1983), Gass and Thompson (1980), Oren (1981), Sargent (1984, 1988), Schruben (1980), and Zeigler (1976).

The Swedish Nuclear Power Inspectorate, SKI, initiated and completed three international cooperation projects to increase the understanding and credibility of models describing groundwater flow and radionuclide transport. The INTRACOIN project is the first of these, and it focused on verification and validation of transport models. The HYDROCOIN study was the second study and represented an international cooperative project for testing groundwater-modeling strategies for performance assessment of nuclear waste disposal. The SKI initiated the study in 1984, and the technical work was finalized in 1987 (Swedish Nuclear Power Inspectorate, 1987). The participating organizations were regulatory authorities as well as implementing organizations in 10 countries. The study was devoted to testing of groundwater flow models and was performed at three levels: computer code verification, model validation, and sensitivity/uncertainty analysis.

Based upon lessons learned from INTRACOIN and HYDROCOIN, international consensus grew prior to and during the GEOVAL Symposium in Stockholm in April 1987 to begin a new project dealing with validation of geosphere transport models. This new international cooperative project, named INTRAVAL, began in October 1987. As with the preceding projects, INTRAVAL was organized and managed by the SKI. The project proposal was based upon a technical proposal developed by an international ad-hoc group from eight selected nuclear waste programs and institutes (Nicholson, 1990).

The INTRAVAL project was established to evaluate the validity of mathematical models for predicting the potential transport of radioactive substances in the geosphere (Swedish Nuclear Power Inspectorate, 1990). The unique aspect of INTRAVAL was the interaction between the experimentalists and modelers simulating the selected test cases for examining model validation issues. The test cases selected consisted of laboratory and field transport experiments and natural analogue studies that incorporate hydrogeologic and geochemical processes relevant to safety assessments of radioactive waste repositories.

These international projects and symposia focused on qualitative aspects of model validation. Very few, if any, touched on quantitative issues. In addition, some of the studies focused on validating a single aspect or observed phenomenon (e.g., matrix diffusion), and none addressed how to validate a predictive, long-term model in a quantitative manner. In the next section, we review some of the studies that tried to evaluate the reliability of predictive models (mostly flow models) that were used for relatively short-term predictions.

4.1 Predictive Reliability and Postaudit

A number of studies have explored the predictive reliability of reasonably calibrated models by a posterior comparison between model predictions and observed data (e.g., Person and Konikow, 1986; Konikow, 1986; Freyberg, 1988). These studies showed that prediction accuracy was moderate. However, the common situation in these studies was that the calibrated model was used to predict system behavior under modified conditions (future predicted system stresses, modified boundary conditions, or different parameter values). In particular, Freyberg (1988) showed that the ability of a calibrated parameter set of a groundwater flow model to reproduce observed data was not an indicator of the ability of that parameter set to predict system response under modified conditions. He reports that good calibration does not necessarily guarantee equally good prediction. Person and Konikow (1986) and Konikow (1986) recalibrated a groundwater flow and solute transport model of an irrigated stream-aquifer system because of the discrepancies between prior predictions of groundwater salinity and observed outcome. They found that the calibration period (covering some seasonal variations in the river-aquifer interaction and irrigation cycles) needed for accurate transport prediction is longer than that required for the flow model predictions. The metric used to judge the prediction accuracy of the model was the spatially averaged groundwater level for the flow model and the spatially averaged groundwater salinity for the solute transport model.

A model postaudit is defined as a comparison of a model's predictions to the actual conditions of an aquifer as a result of the change in conditions (Brown, 1996). Then, "if the model's prediction was accurate, the model is validated for that particular site (Anderson and Woessner 1992b, p. 9)." Anderson and Woessner (1992a) reviewed five postaudits of modeling studies in which the models did not accurately predict the future behavior of the modeled system. These five studies include the studies by Konikow (1986) and Person and Konikow (1986) that we discussed earlier. The other three studies are summarized here. Alley and Emery (1986) examined predictions of 1982 water-level declines and stream flow depletions for the Blue River Basin, Nebraska, made in 1965 using an electric analog model. The postaudit showed that the model underestimated the depletion of the stream flow and overestimated the decline of the groundwater levels. Reanalyzing the model structure, Alley and Emery (1986) concluded that the error in the prediction was a result of uncertainty in the conceptual model of the Blue River Basin.

The next study reviewed by Anderson and Woessner (1992a) is a postaudit study (Lewis and Goldstein, 1982) of a two-dimensional groundwater flow and solute transport model that was developed and calibrated by Robertson (1974). The flow model was calibrated to an assumed steady-state flow field and the transport model was calibrated to observed concentrations of chloride in groundwater in 1958 and 1969. Robertson (1974) then used the calibrated model to predict chloride and tritium concentrations in 1980. Through the postaudit study, Lewis and Goldstein (1982) found that the contaminant plumes predicted by the model extended farther down gradient than the actual plumes and attributed this deviation to the conservative worst-case assumptions in the model input, the simplicity of the conceptual model, and the inaccurate estimate of subsequent waste discharge and aquifer recharge conditions. The original model of Robertson (1974) viewed the aquifer as a continuous porous medium, and it is likely that the flow in this aquifer would be better approximated using a dual-porosity model that includes fracture flow as well as matrix diffusion (Anderson and Woessner, 1992a).

The final postaudit study reviewed by Anderson and Woessner (1992a) is the study of Flavelle *et al.* (1991) that simulates the release of hydrogen ions from a tailings pile situated in glaciofluvial deposits in Ontario, Canada. The flow model of that study was calibrated to measured heads in 1989 within the inner part of the plume where pH was less than 4.8. The solute transport model was calibrated by matching plume position to observed position in 1983 and 1984 through varying the distribution coefficient. The calibrated model was then used to predict the plume distribution in 1989. Data collected in 1989 showed that the model accurately predicted the pH values in the inner core of the plume but not at the outer edges. Flavelle *et al.* (1991) concluded that even though their site is one of the most thoroughly studied uranium tailings sites in Canada, the data were not complete enough for a successful model validation.

Weaver *et al.* (1996) performed a postaudit on two groundwater flow models that were used to design a well array for a groundwater capture and containment system installed along the boundary of a manufacturing facility. The first model was an analytical model for which the postaudit indicated that the performance of the initial system design provided by this model did not meet expectations. This led to using a numerical model to design an enhanced system, for which a detailed postaudit could not be performed, as the system was in place for a short period of time. However, a cursory review of the numerical model results versus observed conditions was performed. The results of the postaudit indicated that the deviations of models' predictions from actual water levels could be mainly attributed to changes in system conditions (pumping rates, variations in well efficiencies, and limitations on total available drawdown) and aquifer heterogeneity.

An interesting discussion related to the postaudit concept is presented by Brown (1996). The previous studies all focused on evaluating the model and conducting the postaudit long after the model had been accepted and used for decision making. So, although the postaudit may enable the modeler to improve the model and benefit from the knowledge gained by the new field data, the improvement can only take place after actions have been taken that were based upon the prediction. Therefore, the postaudit is not something that helps a model withstand attempts at invalidation prior to decision making (Brown, 1996). An alternative to this type of model postaudit is the field postaudit that can be performed after the prediction but before the final decision is made based on the prediction. If some test of a modeling prediction is required prior to decision making, a field audit will provide information of a direct and relevant nature to evaluate the adequacy of a model's prediction. This type of evaluation is what model sponsors and regulators usually call model validation.

4.2 Review of Proposed Validation Strategies

In the context of performance assessment of high-level radioactive waste repositories, Davis and Goodrich (1990) and Davis *et al.* (1991) propose a strategy that focuses on demonstration of model adequacy in representing the real system, given pertinent regulatory requirements, rather than on proving absolute correctness of the model from the purely scientific point of view. In proposing this strategy, Davis *et al.* (1991) take into account the following seven issues: (1) models of performance assessment can never be validated, (2) validation is aimed at building confidence in the model rather than providing a "validated model," (3) model validation implies comparison to reality, but compliance with the scope of regulatory requirements is the overall objective, (4) comparisons to reality should consist of comparing the model results to laboratory and field experiments, natural analogues, and site-specific information, (5) these comparisons will only answer the null hypothesis that the model is invalid,

with the rejection of this hypothesis building confidence into the model, (6) the validation process should consider all plausible conceptual models, and (7) in comparing model predictions to experimental data, a distinction should be made between testing the model structure and testing model input.

The proposed strategy in the above mentioned two studies consists of ten steps: (1) define a validation issue, (2) develop a conceptual model or models, (3) develop a mathematical model, (4) identify and/or design an experiment that addresses the validation issue, (5) define performance measures to be used for model comparisons, (6) quantify the uncertainty associated with the input data and the data available for comparison with the model output, (7) define the acceptance criteria or acceptable model error based on regulatory requirements and data uncertainty, (8) simulate the experiment, (9) perform the experiment in the laboratory or field, and (10) evaluate model results based on the acceptance criteria.

For validating transport models for use in repository performance assessment, Jackson *et al.* (1990, 1992) propose a methodology that includes the following 12 steps: (1) review models, (2) review data, (3) calibrate a specific model, (4) define acceptability of the model with regard to its intended purpose, (5) predict and test, (6) compare with alternative models, (7) analyze discrepancies, (8) assess parameters, (9) present study for review, (10) consider implications, (11) suggest improved experiments, and (12) review consistency.

Along similar lines, Tsang (1987) pointed out the need to differentiate between model structure (geometric structure, geologic units, heterogeneity, etc.) and model processes (dispersion, advection, matrix diffusion, colloidal transport, etc.). He indicates that failure in matching modeling results with field data could be due to errors in the identified model processes and/or model structure. Furthermore, Tsang (1987) makes the distinction that model processes can be validated generically, but model structure validation is a site-specific task. Similarly, Ababou *et al.* (1992) distinguish between testing procedures aimed at checking the internal consistency of complex numerical models and 'groundtruth' experiments, which aim at overall assessment of the model applied to a particular field site. In addition to validating model structure and model processes, Tsang (1987) also proposes the urgent need to validate the procedures for processes and structure identification and the procedures for simplification and conceptualization. Tsang (1989, 1991) reports the need to validate every step of the modeling process in an iterative manner for models that are used for long-term predictions with emphasis on adding an element to the modeling process that can be used to suggest what further measurements are needed to improve the confidence level in the model predictions (e.g., Data Decision Analysis, Pohll and Mihevc [2000]). Tsang (1991) also emphasizes the need for advancing scientific knowledge in related fields, for multiple assessment groups independently studying the same site, and for presenting the modeling efforts in open literature for public scrutiny and evaluation by the scientific community.

Voss (1990) divides the methodology developed for use within the DOE Civilian Radioactive Waste Management Program into the following three general stages: (1) maintaining a record of model development, (2) performing laboratory and field investigations to critically test the model and its premises (e.g., theories, hypotheses, submodels), and (3) carrying out a sequence of formal technical reviews by scientific experts. Voss (1990) also focuses on the importance of approval by the international scientific community regarding model development and model validation. This is to be achieved through comments on reports published in peer-reviewed journals. Voss (1990) also quotes from Kuhn (1982) that publishing the results of

theoretical predictions and measurements in professionally accepted text (textbooks in particular) is in itself establishing reasonable agreement:

"It follows that what scientists seek in numerical tables is not usually 'agreement' at all, but what they often call 'reasonable agreement.' Furthermore, if we now ask for a criterion of 'reasonable agreement,' we are literally forced to look in the tables themselves. Scientific practice exhibits no consistently applied or consistently applicable external criterion. 'Reasonable agreement' varies from one part of science to another, and within any part of science it varies with time…I now conclude that the only possible criterion is the mere fact that they appear, together with the theory from which they are derived, in a professionally accepted text."

Flavelle (1992) proposes a methodology that focuses on the quantitative evaluation of model accuracy when calibrating and validating a model. The method includes performing a regression analysis of predicted values and measured data with the regression coefficient of the regression line interpreted as an empirical indicator of model bias and the standard error interpreted as the uncertainty in the validation. This interpretation provides an initial evaluation of the validation results and the basis for decisions about the usefulness of the model and about the need for more detailed analysis of the validation data. In addition to the simplicity and wide understanding implicit in this analysis, Flavelle (1992) indicates that the approach has the advantage that the validation and calibration statistics can be compared to ascertain if there has been a change in the conditions being simulated, implying that the model does not adequately account for all the important processes.

A linear regression of calculated against measured data provides an initial method to evaluate empirically the quality of the data fit (Falvelle, 1992). Bias in the model and uncertainty in the input and measured data would be expected to affect both the slope of the regression line and the standard error of the regression. Based on this linear regression, one needs to statistically test the assertion that the slope of the regression line is unity and that the intercept of the line is zero. Hypothesis testing can be used for this purpose with the null hypothesis for the slope being H_0: slope = 1, and the alternate hypothesis is H_1 slope ≠ 1. The test statistic is ((slope-1) ÷ standard deviation of the slope). This is to be compared to the critical value of the *t*-distribution at (*n* - 2) degrees of freedom (*n* is the number of data points) and $(1-0\cdot 5\alpha)$ at the α level of significance, $t(n-2,1-0\cdot 5\alpha)$ (Falvelle, 1992). If the absolute value of the test statistic exceeds the critical value, the null hypothesis is rejected. In a similar manner, the null hypothesis of a zero intercept can be examined. Failing to reject both null hypotheses does not mean the model is free of biases, only that this analysis fails to identify any bias (Flavelle, 1992).

Davis and Goodrich (1990) suggested that the deviations of the calculated values from the observations should be examined for trends to identify model bias. The deviations between calculated and observed values correspond to the deviation of observed versus predicted data points from the 45° line on the linear plot. Trends in the set of deviations are what cause the slope of a regression line to vary from unity. Regression analysis has a compelling advantage over analysis of the deviations, as it has been shown that the assumption that the regression residuals are normally distributed is not unreasonable (Draper and Smith, 1981), while the deviations between calculated and observed data may not be normally distributed. Statistical analysis of non-normally distributed data usually requires non-parametric statistical tests, which are more complex than parametric tests used for normally distributed data (Flavelle, 1992).

Luis and McLaughlin (1992) propose a stochastic approach to model validation and apply this approach to a two-dimensional, deterministic, unsaturated flow model for predicting moisture movement during a field experiment carried out near Las Cruces, New Mexico. The model they tried to validate was based on using effective parameter values that were obtained from a large number of soil samples collected before the infiltration experiment at the site. They assume that the model objective was to predict the mean distribution of moisture content over time and space, and postulate that this distribution describes the large-scale flow behavior of most interest in practical applications. The other assumption of their study is that the observations made for the purpose of model validation are small-scale observations collected at sparse points in space and over time.

Luis and McLaughlin (1992) postulate that the differences between predicted and measured moisture content can be attributed to three error sources: (1) measurement errors, which represent the difference between the true values and the small-scale values of moisture content, (2) spatial heterogeneity, which represents the difference between the large-scale trend that the model is intended to predict and the true small-scale values, and (3) model error, which represents the difference between the model prediction and the actual large-scale trend. By expressing measurement residuals in terms of these three components, Luis and McLaughlin (1992) use perturbation analysis and derive the relationship between the measurement residual variance, the actual moisture content variance and the measurement error variance that is only related to the measuring device. This relationship holds only under the assumption that model errors are negligible, and once developed, it can be used to develop statistical tests, which check the hypothesis that the model error is indeed negligible.

Luis and McLaughlin (1992) then applied this approach to the well-instrumented Las Cruces infiltration experiment mentioned earlier. They tried to validate a two-dimensional, numerical model that describes soil properties at the site by a set of spatially uniform effective moisture retention and log hydraulic conductivity parameters, which are inferred from a large set of soil samples collected before the experiment was conducted. The validation approach indicated that this model was able to predict the behavior of the moisture plume at time scales of two years and space scales of 20 meters, but it was not clear that the model would be able to work equally well over longer temporal and spatial scales. The details of this approach are presented in Appendix E and we propose to adapt and use this approach for the current validation plan.

Although this approach provides a quantitative measure to model validation through hypothesis testing, Luis and McLaughlin (1992) caution that this approach should not be blindly applied. In their application to the Las Cruces experiment, which has an unusually extensive set of soil data and validation measurements collected over horizontal and vertical distances of several meters and over time scales of a few years, they could not reach to a conclusion regarding the ability of the model to predict the observed moisture content at later times. In addition, Ababou *et al.* (1992) assert that this approach, although very valuable, is not quite complete since the hypothesis that the model is false remains untested, and the probability of accepting a false model cannot be evaluated by this technique (Chapman *et al.*, 1994). To do this, one would need to postulate another 'complementary' model, or class of models, known to be always true if the model being tested is false. To define and implement such complementary models in an exhaustive fashion is quite a difficult task in the case of spatially distributed phenomena (Ababou *et al.*, 1992).

Mummert (1996) used two validation approaches to evaluate a nitrate percolation model. The first validation method used is a point validation method, where the model accuracy for point predictions is assessed by calculating the coefficient of determination, relative error and standard error. The second validation method used is the statistical validation, whereby Monte Carlo simulations are used to obtain distributions of model predictions. The hypothesis that field data represent "reasonable" samples from the distribution of model predictions is tested by checking whether observed values lie within the five and the 95 percent quantiles of the distribution.

In Appendix VI of the FFACO (2000) for the underground test area (UGTA) at the Nevada Test Site (NTS), the model validation process has a more general and encompassing definition. The ten steps constituting this process are: 1) establishment of the model purpose, 2) development of conceptual model, 3) selection of a computer code, 4) model design, 5) model calibration, 6) sensitivity and uncertainty analyses, 7) model verification, 8) predictive simulations, 9) presentation of model results, and 10) postaudit. This definition implies that to be validated, a model has to go through the entire ten steps using scientifically and technically sound approaches as appropriate for each step. Referring back to the discussion in Section 3.3 on the difference between code verification, model verification and model validation, one can see that the ten-step strategy is better termed as "overall modeling strategy."

What is not clear in the above UGTA strategy is how the five-year proof-of-concept, also prescribed by the FFACO (2000), fits into this ten-step model validation strategy. Also, the details of the postaudit stage and the criteria that govern the pass-fail decision are not obvious. We believe that the proposed validation strategy provides a forum and a structured, systematic approach for making this decision. Therefore, for the CNTA model, the proposed validation strategy can essentially be used to achieve the objectives of step 10 (postaudit) of the UGTA strategy as well as the proof-of-concept analysis through the linkage between validation and long-term monitoring network design. Once the model passes this validation/postaudit stage, the process moves to one of long-term monitoring and stewardship. In these steps, one would use the "final" refined and validated model and predict the output of regulatory interest as well as the output that will drive the long-term monitoring network design. More discussion about the link between the model validation strategy proposed here and the UGTA strategy is presented in Section 6.1.

4.3 Performance Measures, Uncertainty and Acceptance Criteria

The performance measure or the model-produced quantity of interest should be related to a quantity of regulatory interest (Davis *et al.*, 1991). However, regulatory interest in the context of performance assessment models and models of nuclear testing sites usually spans a time scale on the order of 1,000s of years and spatial scales on the order of kilometers. Since experimental analysis designed for model validation studies cannot be extended to these scales, the validation studies must rely on indirect measures for testing the model prediction. This issue raises an important question regarding the extrapolation of model "validity" across multiple time scales. That is, can a model validated (or being declared as not invalid) at certain time (e.g., 50 years after contaminant release) be assumed valid over a time scale of 1,000 years? This issue advocates the importance of long-term monitoring to make sure that model predictions continue to be not invalid as time progresses, which builds increased confidence in model predictions.

While groundwater models cannot be tested over the regulatory scales of interest, certain site-specific factors can be evaluated to gain confidence in using the model to predict flow and transport over these scales. The site-specific data should be used not only as model input data (e.g., hydraulic conductivity), but also as model testing or validation data (e.g., hydraulic heads). Also, a considerable effort should be devoted to justifying the model assumptions as mentioned earlier. Assumptions that cannot be tested or justified by site-specific information have to be tested according to the expert judgment and the acceptance by the scientific community. Anderson and Woessner (1992a) also state that a subjective judgment (based on hydrogeologic expertise and evidence) is always required in deciding whether the mismatch between model predictions and field data is tolerable and that these judgments should be tied to the regulatory purpose of the modeling effort.

Tsang (1987) highlights the importance of the choice of the measurable quantities that are to be used for validation purposes, as there are measurable quantities that are almost impossible to use for model validation (e.g., point and instantaneous concentration data). The averaged solute concentration over a large region and over a period of time is a more relevant quantity for certain purposes such as determining the effectiveness of geological isolation of nuclear or toxic waste (Tsang, 1987). However, depending on the purpose of the model, this comparison may or may not be of importance. Also, more rigorous criteria for upscaling of point concentration or downscaling of model-predicted concentration should be invoked when making such comparisons. Again, this comparison may actually be avoided if one is to only confirm where the plume is, or to evaluate arrival times. Furthermore, the stochastic models provide uncertainty bounds around the best estimate, and field measurements that are properly upscaled can be compared to see whether they fall within or far beyond these uncertainty bounds.

Most of the model components (conceptual model, mathematical model, computer code, and input data) contain some degree of uncertainty due to lack of perfect knowledge about the subsurface conditions no matter how well the system is characterized. Furthermore, experimental results (e.g., field measurements) that are designed for model validation studies contain some errors or uncertainty. The validation tests should consider these sources of uncertainty, which apparently makes it difficult to ascertain whether or not the model results agree with the experimental data; the more uncertain the data are, the more difficult it is to conclude that the model is acceptable (Davis *et al.*, 1991). However, as mentioned before, these uncertainty effects should be viewed in terms of whether or not they affect the quantity of regulatory interest. In some cases, input uncertainty may have minor impact on the resulting regulatory quantity such as the size and location of the water volume having contaminant concentration exceeding a certain threshold (Pohll and Mihevc 2000). These effects should therefore be carefully studied prior to designing the validation study.

Davis *et al.* (1991) discuss some acceptance criteria when validating performance assessment models. To declare that a model is acceptable or adequate for a specific regulatory requirement, the model structure, as well as the model input data, has to be acceptable. Model structure should reflect how the real system behaves. All assumptions inherent in the conceptual model should be justified using site-specific information and field data collected for validation purposes. Accepting the model structure implies that the model results will exhibit a system behavior that is independent of the input data used. That is to say that changing the input data for a structurally accepted model only changes the output results in a quantitative sense but not in a qualitative sense.

To declare that the model input is adequate one has to build confidence in the model over a wide range of experimental conditions. That is, by changing the conditions under which the laboratory or field validation experiment is performed (e.g., different flow or pumping rates), the model predictions can be compared to a wide range of input conditions that will help build confidence in model input. When changing experimental conditions and thus some portions of the input data for the model, the adequate model should predict the experimental results with a reasonable accuracy without changing other input data. If other input data are correlated to those changing conditions, then the model input should reflect this type of correlation to accept the model input and declare the model not invalid.

5. CONSIDERATIONS AND CRITICAL ISSUES

5.1 Reducing the Prediction Uncertainty

Validation of predictive models should provide confidence in the uncertainty band of the results, within which the real outcome will fall (Zuidema, 1994). Understanding the impossibility of completely eliminating uncertainty (Gorokhovski and Nute, 1996), we should develop ways of making groundwater models and the decisions based on them more reliable and effective. The proposed plan focuses on making use of collected validation data to reduce model uncertainty and narrow the range of possible outcomes of stochastic numerical models. This requires iterative implementation of data collection, model evaluation, model refinement, and uncertainty reduction and is particularly important in radionuclide transport models as only small aspects of the transport model results can be tested. In this case, our proposed validation approach would focus on the non-transport elements of the model (e.g., geology, structure, and flow) and use the validation data to refine transport predictions and reduce their uncertainty.

As pointed out by Anderson and Woessner (1992a), a partial validation may be achieved by the demonstration that a good modeling protocol is implemented in the modeling process and by a thorough assessment of model calibration and uncertainty analysis. The use of validation data to reduce prediction uncertainty is thus an important step in the validation stage where refining the model with new data helps build increased confidence in the model. This trial-and-error approach together with the understanding that uncertainty cannot be completely eliminated represents important aspects of the validation approach and should be clearly presented to model sponsors and regulators for their understanding and approval.

5.2 Diversity of Data and Evaluation Tests

As discussed by Ababou *et al.* (1992), the degree to which a single experiment (or a single set of field data) can validate a model depends on the subjective weights, or probability, assigned to that particular experiment. More validation weight can be assigned if the range of aspects covered by the experimental data set is broad enough that the overall character of the model is efficiently put to test. The field data should, therefore, be diverse and cover different aspects of the model. For example, the data should be able to test geologic aspects (e.g., the existence and location of contact between different geologic units), flow model aspects (e.g., head and gradient measurements), and transport or contaminant release aspects (e.g., concentration measurements). Since one of the purposes of the validation task, if not the most important one, is to see if multiple failure and far-field transport of contaminants can at all take place, transport aspects related to some failure scenarios should be tested.

Oreskes *et al.* (1994) postulate that by using as numerous and diverse confirming observations as possible, it is reasonable to conclude that the conceptualization embodied in the

model is not flawed. Therefore, a diversified set of statistical tests and evaluations for the model will provide a structured approach for evaluating the model predictions and building confidence in the decisions based on these predictions. The systematic validation approach we propose here relies on a number of different tests and evaluation techniques that will help guide the decision regarding the model predictions and allow for informed and grounded discussions among the modelers, model sponsors and regulators.

5.3 Submodels

In general, if a single model is divided into two or more submodels, the degree of confidence imparted by evaluating the submodels individually will not be as great as the degree of confidence achieved by evaluating the submodels linked together (Eisenberg *et al.*, 1994). Therefore, it is important to perform additional tests to validate the combined submodels. Site-specific groundwater flow and transport models can be divided in general into three submodels that can be tested individually first and then combined. Figure 1 shows an example of the different submodels of a site-specific model and how they are linked to each other. The figure shows both the conceptual and the numerical submodels.

For the first submodel, a geologic model identifying the different units and how they are structured together within the study domain is conceptualized. The input to the first submodel constitutes all the data types that help identify the geologic units and where they are located (e.g., lithologic data, geophysical logs, resistivity logs). With categorical or qualitative data and using geostatistical tools and conditional simulation, a discretized numerical submodel of the different categories or units can be obtained. Subsequently, one can use the quantitative data available (e.g., hydraulic testing results, packer tests, resistivity logs) to obtain the detailed heterogeneous structure of each individual unit in a quantitative manner. That is, the spatially varying hydraulic properties, namely hydraulic conductivity, can be obtained as an output of this first submodel.

For a general site-specific model and for the special case of the CNTA model, the first submodel can be tested in terms of the existence and location of the different units identified in the conceptual geologic model. Contact between the different units is also an important aspect that can be tested with validation data. For the CNTA model, Pohlmann *et al.* (1999, 2000) identify three geologic units with significant uncertainty associated with the contact between them. Conductivity values assigned to different layers should also be evaluated. This evaluation will focus on reducing uncertainty in the assigned conductivity values by utilizing head measurements and a conditional simulation (or inverse) approach. For example, the sequential self-calibration (SSC) approach (Wen *et al.*, 1996; Gómez-Hernández *et al.*, 1997) can be used for this purpose.

The second major submodel for a general site-specific groundwater model is the flow submodel, where the output of submodel (1) is used as input. A conceptual flow model is then formulated and used in conjunction with this input and boundary conditions and assumptions to derive the numerical flow model and solve the flow equations. This results in identifying the flow pattern in the simulation domain, which is represented by discretized head values and velocity components. This velocity distribution is the output of submodel (2) and is used as input to submodel (3).

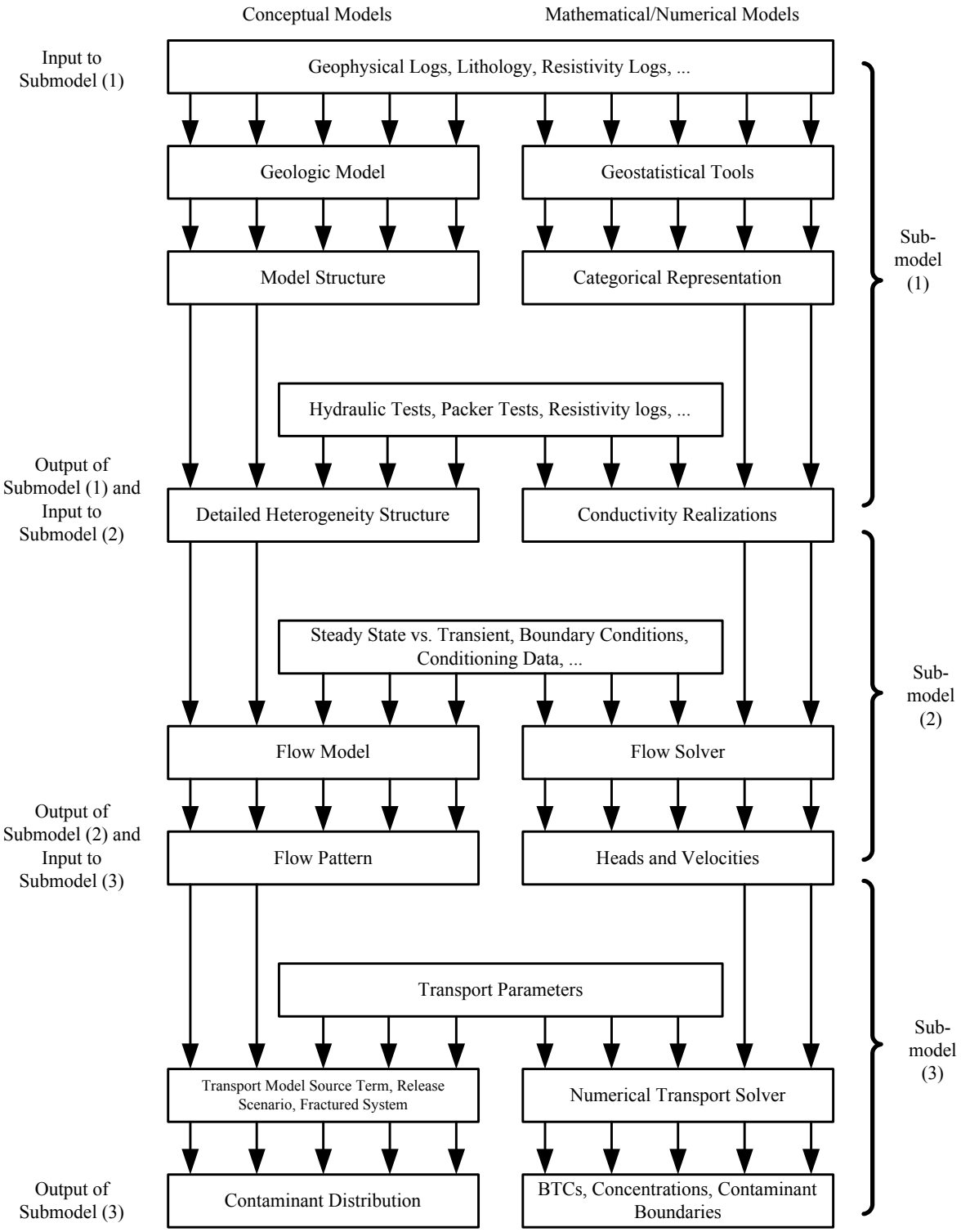

Figure 1. A schematic representation of a general site-specific groundwater flow and transport model showing the conceptual and numerical models and the three main submodels linked together.

The flow pattern at CNTA (and at many other field sites) is fairly complicated (see Pohlmann *et al.*, 2000) and it is crucial to verify the directions of the vertical and lateral head gradients, especially in the vicinity of the contaminant source. Multiple head measurements at different levels can be obtained from a single borehole, and these data will be crucial to testing the flow model and its underlying input data and boundary conditions. In addition to testing the predicted heads themselves, the head data will be used to reduce the heterogeneity uncertainty by using an inverse method such as the SSC approach mentioned above.

In general, the last submodel in a site-specific study is the transport model. The conceptual transport model is identified by determining the source size and location, the release scenarios, and the transport processes encountered during the migration of contaminants. Added to the velocity pattern and boundary conditions, this conceptual model gives rise to the numerical transport model where the transport equations are formulated and solved for the output of concern. This solution yields temporal mass flux breakthrough curves at certain boundaries, spatial-temporal distribution of contaminant concentrations, or contaminant boundaries. Usually, these latter outputs are the target of the entire modeling process when groundwater contamination is the major regulatory concern.

For the CNTA transport model, the release of radionuclides from the test cavity and the movement away from it are just starting (based on a cavity infill time of 30 years for a test conducted in 1968). An important focus of validation of the transport aspects should be verifying whether there are any fast migration channels or failure scenarios that may have been overlooked and would thus lead to migration distances greater than the model predictions. Measurement of tritium concentrations in wells located sufficiently far from the cavity (e.g., beyond the fracturing radius to separate the possibility of fast migration pathways from prompt injection issues) will be important to test the adequacy of the transport model and whether or not the model (within its uncertainty bounds) has covered all the critical transport issues.

After considering the different components and tests described above and linking the calibration analysis to the validation analysis, we arrive at the stage of evaluating the different submodels linked together. The flow of information between the three submodels provides a natural linkage that will enable collective evaluation of the entire model to be conducted in parallel with the individual submodel evaluations.

5.4 Subjective Versus Objective Judgment

Calculated and observed data for both the calibration and validation processes most often are presented graphically, with a subjective interpretation of the quality of the match (Flavelle, 1992). It is generally preferable, however, to use some form of objective analysis to perform model calibration and validation. The objective quality of model calibration is usually described by a goodness-of-fit parameter, which reflects how well the model results match the observed calibration data. The goodness-of-fit is usually used to optimize the calibration of the computer model's adjustable parameters and to serve as a measure by which to compare alternate models. This is an inverse problem, for which the main problem is the non-uniqueness of the solution that gives rise to obtaining different parameter values that yield solutions with similar accuracies (e.g., Poeter and Hill, 1997; Hill *et al.*, 1998; D'Agnese *et al.*, 1999). The most common goodness-of-fit parameter appears to be some form of weighted root-mean-square error, with the error describing the difference between calculated and measured values. Unfortunately, while quantitative evaluation of the quality of model calibration is becoming more common, the

complexity of some of these evaluations makes them unattractive for general use by regulators and decision-makers (Flavelle, 1992). It is, therefore, more appealing to invoke simple goodness-of-fit tests and describe the calibration and validation processes in an objective manner. The validation approach proposed here relies heavily on objective evaluations and a number of statistical measures and tests for evaluating different aspects of the model.

A common form of objective analysis for calibrating and validating simulation models is statistical hypothesis testing (Balci and Sargent, 1981, 1982). Two types of errors can exist in hypothesis testing and may lead to wrong decisions if testing results are used for decision-making. The first type of these two is called type I error, which results from rejecting a hypothesis while in fact it is a correct one (e.g., rejecting the validity of a valid model). The second is referred to as type II error, which results from accepting a false hypothesis (e.g., accepting the validity of an invalid model). Despite these errors, we propose to use this form of objective analysis in addition to some other goodness-of-fit tests to evaluate the quality of the comparison between model predictions and measurements for both calibration and validation. More background details are given in the Appendices regarding goodness-of-fit measures and hypothesis testing.

McCombie and McKinley (1993) argue that the decision about how much effort must go into the validation process before the model can be considered to be valid is necessarily subjective and very dependent on the complexity of the system and on the objective of using the model in the first place. They further recommend that the subjective aspect of assessing if a model is good enough be included in the term "validation." This argument and the above discussion highlight the fact that neither purely objective judgment nor purely subjective judgment can be used in the validation process. In other words, each of the objective and subjective judgment components is a necessary component for the model validation process, but is not a sufficient tool. They complement each other, and model builders, model users, and regulators should come to an agreement that objective judgment will always be complemented with subjective judgment and hydrogeologic expertise.

5.5 Validation Cost and Confidence in the Model

The cost of obtaining data and performing the analysis for model validation should be considered in designing any validation plan. As shown in Figure 2, adapted from Sargent (1990), there is a limit beyond which increased investment in model validation efforts (both data collection and analysis) does not significantly increase confidence in the model and adds little value to the end user (Sargent, 1990). Therefore, the model validation process requires consent between concerned parties regarding the level of confidence required for the model to be validated, keeping an eye on the cost that is needed to achieve this confidence level. The proposed validation approach discussed in the next section has a number of decision points, for which the cost of the data collection and analysis comes into play in making these decisions.

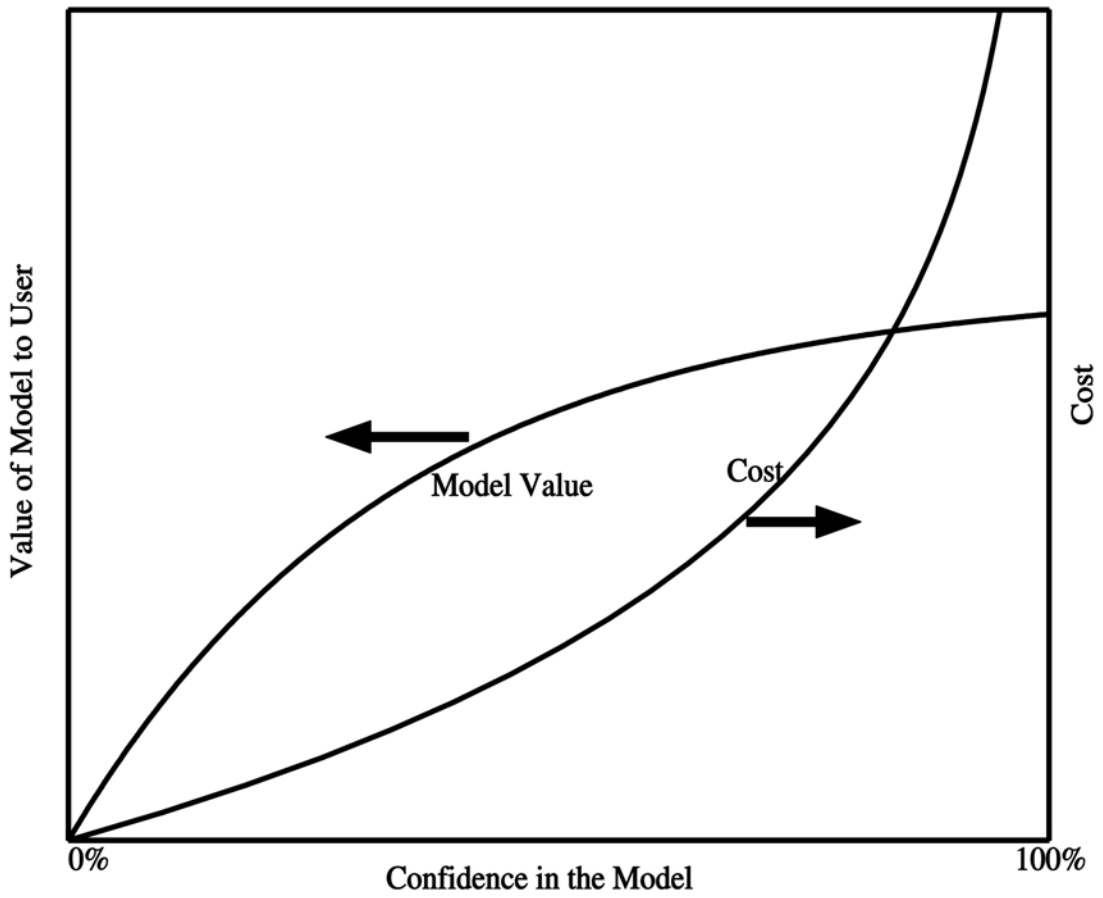

Figure 2. The change in model value and in the cost of investing in model development and validation as a function of the desired confidence level in the model (adapted from Sargent, 1990).

6. PROPOSED VALIDATION APPROACH

6.1 General

The effectiveness of a validation strategy, i.e., its ability to discriminate between good and bad model assumptions, depends on the type of available data and how the data are used to challenge these assumptions (Mroczkowski *et al.*, 1997). These authors argue that validation using multi-response data is a considerably more powerful strategy than traditional split-sample testing (where a record of historical data is split into calibration and validation samples). They, however, base their argument on the validation of conceptual catchment models, where large historical records exist for the parameters studied. One would expect that the use of multi-response data would also be much more powerful in validating a subsurface flow and transport model than using a single type of response data. Our proposed approach to model validation relies on using both multi-response data and diverse statistical tests and analyses to evaluate model performance. By doing so, one can build confidence in the model predictions and guide the field activities for collecting the data needed for the long-term monitoring of the site.

To determine the accuracy of the model and its adequacy, one should consider the types of validation tests, the number of validation tests, the degree of agreement between model and the validation tests and the conformity between model descriptions and site-specific information (Davis *et al.*, 1991). These authors emphasize the necessity for rigorous development of the validation process and the importance of providing regulators with validation information that is as inclusive as possible and follows a logical systematic approach. The approach we propose here relies on numerous validation tests and evaluations and follows a systematic step-by-step approach as will be discussed in the next section. This systematic approach is particularly crucial when it comes to validating stochastic numerical models that rely on Monte Carlo simulation techniques, where multiple realizations within this stochastic framework need to be analyzed and evaluated in a systematic manner.

A unique aspect of the CNTA validation plan is that it is the first attempt to validate a stochastic model that explicitly accounts for spatial variability in conductivity and parametric uncertainty. The literature review and the discussions presented in Hassan (2002) and briefly summarized in section 3 make it clear that even the simplest deterministic subsurface model is very difficult to evaluate. The proposed plan accounts for the stochastic nature of the model and attempts to reduce the realm of possibilities given by the large number of realizations considered in the Monte Carlo analysis.

As indicated in the previous sections, there are currently no algorithms or procedures available to identify specific validation techniques or statistical tests that can be used in a complete manner in the validation process. In addition to the data scarcity and other challenges facing model validation, the CNTA transport model indicates that the nuclear test cavity infill is about to be complete, which means that transport migration away from the test cavity has not begun (Pohlmann *et al.*, 2000). Despite these challenges, we need to build confidence that model-based decisions will not result in unacceptable risks to present or future populations or in degradation of the natural environment (Konikow and Bredehoeft, 1992). Building confidence in the models used to support closure of sites is the requirement for validation; developing a validation process that allows regulatory closure of sites with significant groundwater contamination should, therefore, be the ultimate goal of any validation strategy. We propose a systematic approach for validating the CNTA groundwater model in a manner consistent with the ultimate use of the model and the regulatory requirements. The rigor of the proposed approach stems from its simplicity, comprehensiveness, and coverage of many aspects of the model rather than its mathematical complexity.

Many of the tests that we propose to use in the validation approach and their underlying principles are familiar. The power of these tests and the power of the integrated validation approach stem not from their innovation but from their rigor and completeness. We are not developing new theories or statistical analyses, but rather putting together a number of available tools in an integrated manner to evaluate groundwater models that are used for decision making. Together, these tests and the proposed systematic validation approach provide a structured approach for analyzing all the key issues and components of a site-specific groundwater model in the hope of building confidence in the decisions based on the model predictions. Individual decisions throughout the validation stage will still be difficult, often requiring subjective judgment and some trade-offs, but using the structured, systematic validation approach we propose here will help guide the decision and make the debate among involved parties more rational.

Our philosophy in developing and advocating this validation approach relies on a forward-looking perspective. That is, by carrying the groundwater modeling process one step further beyond the small iterative loop of characterization-calibration-modeling-prediction and back to characterization to reduce uncertainty, one can learn a great deal about the site and the model together. Unfortunately, no matter how many times the iterative process is repeated, there will always be a level of uncertainty about the results of these studies and whether they represent reality or not. Without a way to exit the loop of characterization, conceptualization, calibration, modeling, and back to characterization, resources may be allocated to efforts and studies that do not ultimately resolve the problem of concern. The flow chart shown in Figure 3 schematically represents this loop (thin-lined loop) and proposes a logical way to exit this loop. This occurs through the groundwater flow and contaminant transport validation process (the outer, bold-lined loop in Figure 3), which develops a systematic method of determining when adequate confidence in the groundwater model has been achieved and long-term monitoring should begin. It is possible, of course, that model deficiencies can drive the process back to the inner loop of characterization, but this would only occur after analysis of validation and monitoring results over time.

Figure 3 also shows the linkage between the proposed strategy and the ten-step UGTA model validation strategy. Steps 1 through 9 of the UGTA model validation strategy belong to the development stage that is represented in the figure by the path from characterization to the results circles within the closed, thin-lined loop. Step 10 in the UGTA strategy (the postaudit) is highlighted by the bold lines in Figure 3, and it represents the model validation conceptualization as perceived in this study. It is important to notice that the five-year proof-of-concept monitoring network development that is required in the FFACO (2000) can start once the development loop is exited and can be performed simultaneously with the model validation analyses, monitoring network development and the postaudit. As stated in the FFACO (2000), measurement of field parameters through this proof-of-concept monitoring network will be used to demonstrate that the model is capable of making reasonable predictions that fall within an acceptable level of confidence. This is exactly what the closed, bold-lined loop in Figure 3 is designed to provide. When the initial monitoring network is installed, data collected, and evaluation tests performed for evaluating and validating the model, the question will arise as to how the model predictions compare to the collected field measurements. If the results indicate major model deficiencies, the process will be driven back to UGTA step 2 (the leftmost bold, upward arrow in Figure 3). Steps 2 through 9 of the UGTA strategy will be repeated and this repetition is considered as part of the model postaudit or model validation stage. If our confidence-building, bold-lined loop indicates reasonable model performance, but more confidence building is still needed, the process is driven back to the simultaneous stages of selecting validation targets and developing (or augmenting) the proof-of-concept monitoring network. Here, new well locations may need to be determined to augment the initial monitoring network. This will provide additional wells for growing the five-year proof-of-concept monitoring network, and this iterative process continues until sufficient confidence is built in the model formulation and predictions.

Once UGTA step 10 (model postaudit), the proof-of-concept stage, and confidence-building loop (model validation process) have been completed successfully and the model is deemed validated, the design will start on the long-term monitoring network that will augment existing wells so as to provide sufficient surveillance for the site. Using the long-term monitoring data to re-evaluate the model over time is considered as a continuous model validation and postaudit process and is necessary for the long-time period of concern at these nuclear testing

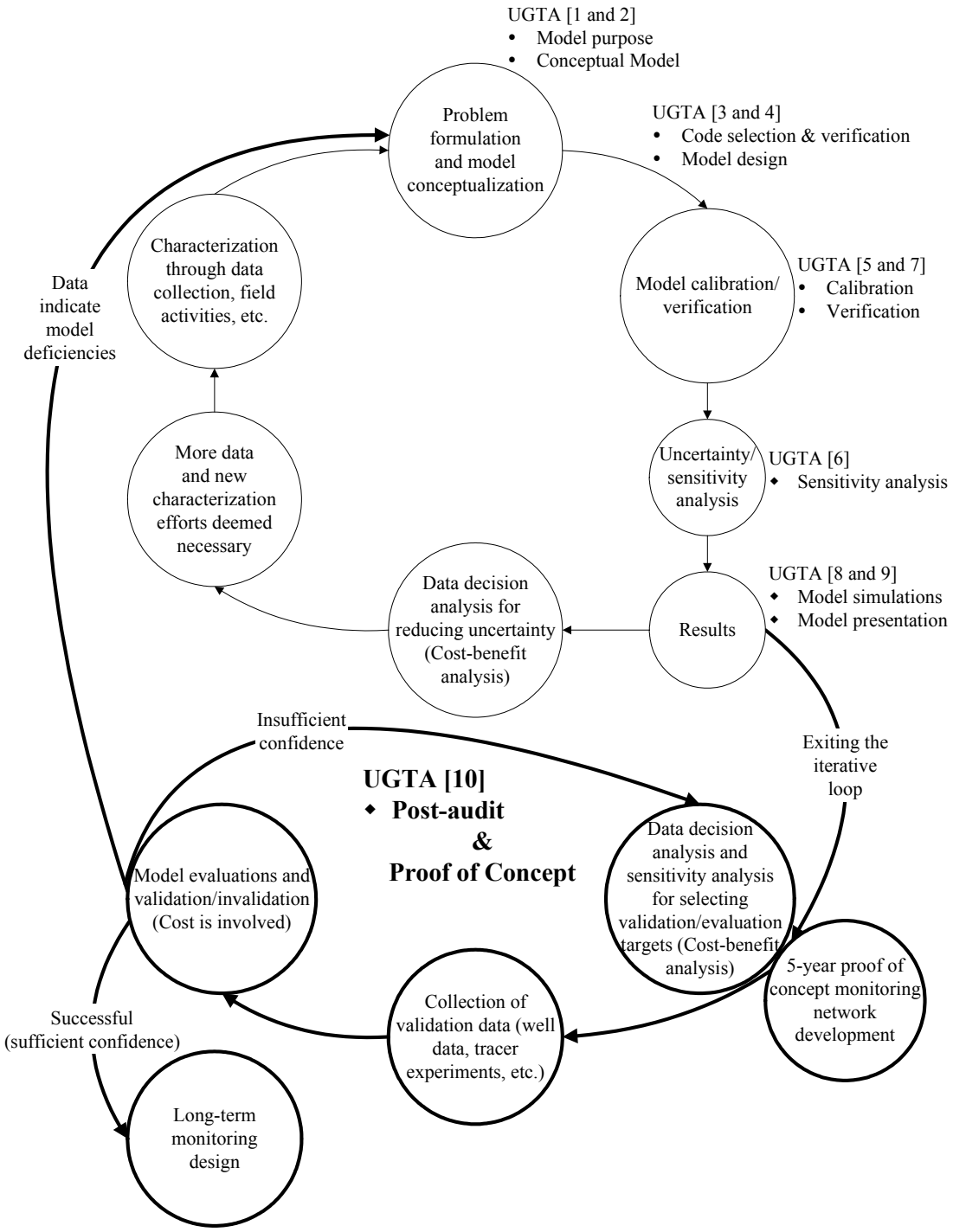

Figure 3. A schematic representation of the characterization-calibration-modeling-characterization loop (thin-lined loop) and the way to evaluate the process to either exit the loop and start long-term monitoring or continue for better characterization and modeling. Also shown is how this process relates to the UGTA ten-step model validation strategy as outlined in Appendix VI of the FFACO (2000).

sites. Again, the how-to steps related to the UGTA model postaudit stage and the link to the proof-of-concept monitoring are not mentioned in any quantitative manner in FFACO (2000). What is provided in Figure 4 and discussed in Section 6.2 is a detailed approach for performing this process in the case of stochastic numerical models that rely on Monte Carlo simulations, which is the case for CNTA and other UGTA and offsites studies.

It is important to note that previous studies that dealt with groundwater model validation (e.g., Tsang, 1991) focused only on the small, iterative loop (thin-lined loop) shown in Figure 3. For example, Tsang (1991, Table 1) asks the question of whether the "evaluation of the results" indicates that uncertainty is too large or results with estimated uncertainty are good enough. This is essentially equivalent to the data decision analysis step in the thin-lined loop in Figure 3. Also, previous studies did not explicitly consider the stochastic nature in a Monte Carlo fashion as is considered here for the CNTA stochastic model. Furthermore, the quantitative aspects were absent in previous studies, whereas the proposed approach contains many quantitative tools such as goodness-of-fit measures, hypothesis testing, and regression analysis in evaluating model results.

Going forward with the validation analysis will allow one to say "maybe" the model is good enough, while staying in the small iterative thin-lined loop of Figure 3 will not allow any judgment regarding the model. There will never be enough facts or data to eliminate all the uncertainty or to make a decision based solely on those facts. It is, therefore, better to move forward in the face of uncertainty and make decisions regarding the model conformity with regulatory requirements, and then evaluate these decisions periodically over time.

6.2 Proposed Step-by-Step Procedure for Model Validation

To start, the steps to carry out the proposed model validation and the refinement of the model predictions based on the collected validation data are listed. Detailed theoretical background and descriptions of the different steps are presented in the Appendices. The proposed steps, shown in the flow chart of Figure 4, are as follows:

Step 1: Identify the data needed for validation, the number and location of the wells, and the type of laboratory or field experiments needed. The well locations can be determined based on the existing model and should favor locations likely to encounter fast migration pathways. This step will mark the beginning of the five-year proof-of-concept monitoring network development stage. The well locations will be determined using a monitoring network design approach to provide measurements that will be used to show whether the model is capable of making reasonable predictions that fall within an acceptable level of confidence. There are additional factors guiding well location, which are determined by the site conditions and the nature of contamination. For example, for the CNTA model, the first consideration is that wells should be located far enough outside the fractured radius of the zone impacted by the nuclear test to avoid confusing prompt injection of radionuclides from the blast with radionuclide migration. Second, the wells should be located around the cavity in such orientation to obtain the most benefit from them in validating the model and refining it. The layout of the wells should be designed to enable a verification of lateral and vertical head gradients and flow directions around the cavity area. Other factors such as safety issues associated with radioactive contamination and the cost of drilling and collecting data have to be considered. Sequencing of data collection is also

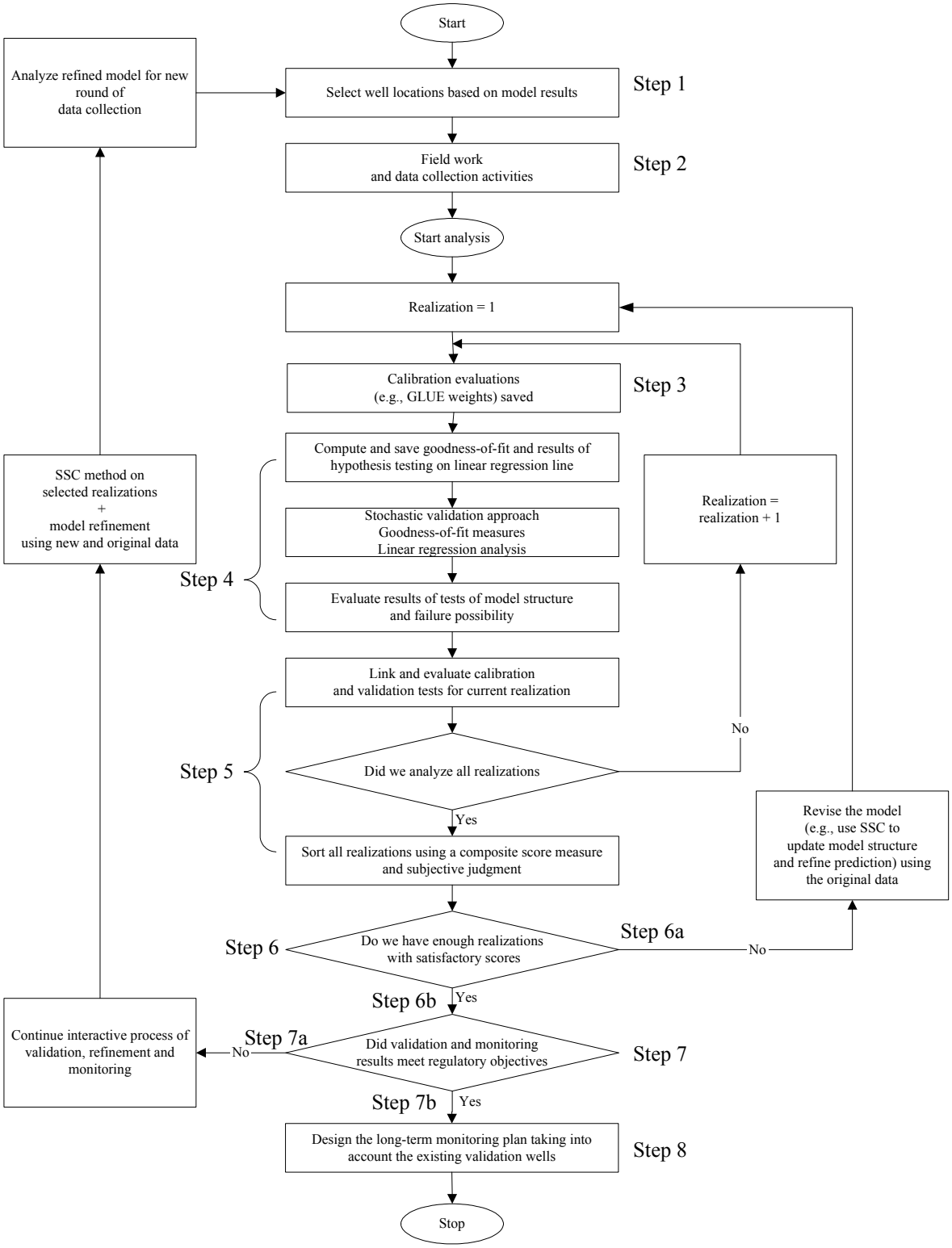

Figure 4. A flow chart showing the proposed validation approach and the associated iterative refinement loops.

important. Though it may be more practical and cost-efficient to drill the wells simultaneously, drilling one well at a time and collecting all possible data from it and testing the model to determine the next field activity may be a better approach. Again, these choices depend on the problem at hand and need a consensus among model developers and model users.

Step 2: Carry out the fieldwork to install the wells and obtain the largest amount of data possible from the wells. The data should include geophysical logging, resistivity logs, head measurements, concentrations (e.g., checking for tritium), and any other information (e.g., temperature logs, conductivity measurements) that could be used to test the model structure, input, or output. The philosophy here is that the major portion of cost in the case of deep groundwater contamination (e.g., nuclear testing sites) is incurred during drilling the wells. Therefore, it makes a good investment to collect as much data as possible from these wells, because the extra data collection cost (in the short-run) is going to be marginal in comparison to the drilling cost.

Step 3: Evaluate the calibration accuracy for each individual realization using different goodness-of-fit measures in addition to the generalized likelihood uncertainty estimator (GLUE) (Freer *et al.*, 1996; Franks and Beven, 1997; Pohlmann *et al.*, 2001). This assumes that the original model calibration was qualitative in nature (which is a common case) and was done to minimize the deviation between model prediction and observed calibration data based mainly on visual inspection. A more detailed discussion of the GLUE analysis is presented in Appendix A. Other tools such as linear regression analysis, goodness-of-fit tests and hypothesis testing can be used to provide additional objective means to evaluate the relative strength of each realization in terms of reproducing the field calibration data. That relative strength will be linked later to the ability of individual realizations to match the validation data. Appendix B presents some description and discussion of goodness-of-fit measures and Appendix C reviews some linear regression analyses and benefits in model calibration and validation. Also, the use of hypothesis testing for model evaluation is briefly described in Appendix D.

Step 4: Perform the different validation tests that will help evaluate the different submodels and components of the model. A promising stochastic validation approach was proposed by Luis and McLaughlin (1992) and was applied to a two-dimensional, deterministic, unsaturated flow model for predicting moisture movement during a field experiment carried out near Las Cruces, New Mexico. A detailed description of this approach is summarized from Luis and McLaughlin (1992) and presented in Appendix E. This approach can be adapted and used to test the flow model output (heads) under saturated conditions. Other objective tests (e.g., goodness-of-fit tests) can be used for the heads to complement this stochastic approach that is based on hypothesis testing. Similar tests will be performed to test model structure and or input depending on the type of data to be obtained in the field. Some data will be used to check the occurrence or lack thereof of failure scenarios (e.g., at CNTA one needs to check if tritium exists much farther from the cavity than is predicted by any realization of the stochastic model). The philosophy here is to test each individual realization with as many diverse tests (in terms of the statistical nature of the test and the tested aspect of the model) as possible and have a quantitative measure of the adequacy of each realization in capturing the main features of the modeled system.

Step 5: Link the results of the calibration accuracy evaluations and the validation tests for all realizations and sort the realizations in terms of their adequacy and closeness to the field data. A subjective element may be invoked in this sorting based on expert judgments and

hydrogeologic understanding. The objective here is to filter out the realizations that show a major deviation or inadequacy in many of the tested aspects and focus on those that "passed" the majority of the tests and evaluations. By doing so, the range of output uncertainty is reduced and the subsequent effort can be focused on the most representative realizations/scenarios. To continue reducing the uncertainty level, a refinement of the conductivity distribution can be made using the SSC method mentioned earlier and described in Appendix F. In this method, head (and concentration) measurements can be used to condition the generation of the conductivity field in such a way that the uncertainty in the conductivity heterogeneity pattern around each measurement location is reduced. This updating of the conductivity distribution can be done for each of the original conductivity realizations that were retained in the analysis.

Step 6: The results of step 5 will determine the forward path and guide the decision as to whether there is a sufficient number of realizations that attained a satisfactory high score (thus building confidence in the original model) and are considered sufficient for further analysis or whether this number of realizations is not sufficient in comparison to the realizations with low scores indicating that the original model needs major revisions.

6a. If the number of realizations with low scores is very large compared to the total number of model realizations, it is an indication that the model has a major deficiency or conceptual problem or that the input is not correct. In this case, the conceptual model should be revised and model structure updated based only on the original calibration data if possible. This means that the validation data should not be used and in essence should be forgotten. This is done to avoid new validation data collection at this stage when the previous analyses indicate that the model is inadequate as is. If this is difficult, however, a compromise solution could be to split the validation data set and use part of it in the model refinement process and save the other part for the next round of validation tests and analyses. The possibility also exists that after the model is refined, new wells at different locations will be needed (e.g., if the analysis indicates a shift in the flow direction such that the initial monitoring network will not be optimally located for collecting the relevant data). In this case, the five-year proof-of-concept monitoring network will be modified and used for the new round of validation data collection. This iterative process, when eventually completed, will in essence provide the evidence that the monitoring network is doing what it is supposed to do, which is the main purpose of the five-year proof-of-concept stage of the entire process.

6b. If the number of realizations with high scores is found sufficient, this indicates that the model does not have any major deficiencies or conceptual problems and one can move forward to step number 7.

Step 7: Once the rightmost loop in Figure 4 is completed successfully and a sufficient number of the model realizations show acceptable performance (this is judgmental and should be based on the hydrologic expertise and judgment of the researchers involved), the model sponsors and regulators in collaboration with the model developers have to answer the last question in Figure 4. This question will determine whether the validation results meet the regulatory objectives or not. Anderson and Woessner (1992a) suggested that regulators should be content with some degree of partial validation and should further shift the focus from demands for validation to demands for good modeling protocol that includes a complete description of model design, a thorough assessment of model calibration and an uncertainty analysis. It is important to recognize that the data collected represent both validation and monitoring data. The five-year proof-of-concept monitoring network development will essentially occur at the beginning of the

validation process (postaudit in FFACO's terms) and the data collected will be used to "demonstrate that the model is capable of making reasonable predictions that fall within an acceptable level of confidence." Thus the question posed at this stage is whether the designed five-year proof-of-concept monitoring network provides sufficient surveillance for the site and whether the collected data and the resulting evaluation tests provide sufficient evidence about the fidelity of the model.

7a. If the answer to the question posed is no and there is a need to collect more data for more confidence building in the model or that the monitoring network needs to be modified, then the left-hand-side loop in Figure 4 gives rise to a new iteration of model refinement, new well placement, data collection and re-evaluation. In this case, all available data become calibration data and new data will need to be collected for validation, probably from new wells. Steps 1 to 6 are repeated with the data to be collected determined based on the analysis of the refined model. It is thus better to benefit from the validation data and refine the model using the representative realizations before proceeding to the new round of data collection. The new wells for this round should be selected to serve two purposes: 1) sources for the new validation data and 2) location targets for the long-term monitoring of the site.

7b. If the answer to the question posed is yes, validation is deemed sufficient and the model is considered adequate or robust and we then proceed to step 8.

Step 8: Design a long-term monitoring plan. This includes setting and clarifying the objectives of the monitoring, designing the monitoring networks, determining the frequency of sampling, where, when and what to sample, etc.

The above steps outline the general approach we propose for validating stochastic numerical groundwater models that rely on Monte Carlo simulations. Figure 4 shows a flow chart that summarizes these steps and the iterative process that needs to be implemented for building confidence in groundwater predictive models and moving toward the long-term monitoring and closure of contaminated sites. The approach is general in nature and the application to the CNTA model will be the first attempt to validate a stochastic model for a nuclear testing site to the best of our knowledge. The iterative nature of the proposed approach is one of its greatest strength. Numerical groundwater models, and in particular stochastic models, are very complex and modifying or changing any aspect of the model may produce unanticipated consequences in a different aspect of the model. To get the best outcome of the validation process, one needs to both consider the different details separately and take the broader view of the entire model while working step-by-step through the different decisions and trade-offs.

It can be seen and expected that the process of validating a site-specific groundwater model is not an easy one. Throughout the structured process described above, we may wonder whether there is any way to know and confirm that we are on the right track. It is our belief that the way to this confirmation is the cumulative knowledge gained from the different stages of the validation process. That is, a set of independent tests and evaluations will provide a great knowledge about the model performance and their results will provide some incremental, but additive, pieces of information that will be of superior importance. While there are no guarantees of success (attaining a conclusive outcome about model performance), the combined presence of these different results and evaluations sharply improves the odds that one can make a good decision about the model performance.

As mentioned earlier, an attempt is made during the development of this approach to honor most of the critical issues raised in previous groundwater model validation studies and discussions. For example, Konikow (1986) states that models should be considered as dynamic representations of nature, subject to further refinements and improvements. As new data become available (e.g., through new wells), model predictions can be evaluated, validated or invalidated, and then modified if necessary. This dynamic loop is considered in the proposed validation approach outlined above. Also, Tsang (1989, 1991) argues that it is important to validate every step of the modeling process in an iterative manner for models that are used for long-term predictions with emphasis on adding an element to the modeling process that can be used to suggest what further measurements are needed to improve the confidence level in the model predictions. Along similar lines, Anderson and Woessner (1992a) state that conceptual models need periodic improvements through data collection and a trial-and-error process of evaluation over many years.

For completeness, we present all the necessary background regarding the proposed tests and techniques in the Appendices. This background information is simply compiled from the different studies cited in the Appendices and is aimed at clarifying the proposed approach and summarizing the necessary tools for this approach in one document. It should be mentioned, however, that these tools are just examples of many statistical tools and techniques that can be used to achieve the same goals and meet the same objectives. Examples are given here in the hope that other techniques and approaches are developed that will enhance the proposed validation approach and make it more practical and appealing to model sponsors and regulators.

7. CONCLUDING REMARKS

The challenge of validating numerical models, especially subsurface models, arises not only from the technical and scientific difficulty, but also from the lack of widely accepted definition of the term itself and the purpose of the process of validation. This report is an attempt to summarize the different validation perspectives and definitions, to analyze their merits, and to propose a model validation plan for evaluating the CNTA model. Important definitions and the distinctions that have to be made when dealing with the terms “calibration,” “verification,” and “validation” are highlighted. A review is presented of studies that deal with groundwater model validation and propose certain validation strategies. Common to most, if not all, of these studies is the fact that no quantitative objective tools were provided in an integrated manner in any of the proposed approaches, making them difficult to adapt and use in different situations. It is also common among these studies that the general consensus of the hydrogeologic community is that absolute validity (accurate or exact representation of reality) is not even a theoretical possibility and is definitely not a regulatory requirement. Confidence building in the modeling process and in the subsequent evaluation and validation procedure is viewed as the best way to achieve model validation objectives and acquire acceptance of the regulators and the public.

Building on this review, a groundwater model validation strategy should take into account a number of important issues that were recognized as being important to the process in many of the reviewed studies. These issues include reducing prediction uncertainty, diversity of data and evaluation tests, relying on objective measures whenever possible and also capitalizing on subjective judgment and hydrogeologic insights, testing the different submodels individually and in connection to one another, and recognizing that the cost element of the validation process will play a significant role in making many of the decisions throughout the process. Considering these issues and the fact that the confidence building process in model prediction is a long-term

and iterative process, a systematic approach for the general case of a stochastic numerical model has been developed and is proposed for the CNTA model evaluation. One of the main outcomes of this study is an integrated validation approach that relies on an iterative calibration-modeling-monitoring-evaluation-refinement cycle, which would eventually increase confidence in CNTA model predictions and reduce the uncertainty level associated with these predictions. The methodology will be fully developed, tested, and enhanced during the implementation and application to the CNTA groundwater flow and transport model.

Opponents of the use of the term "model validation" postulate that the term is misleading to the public because it conveys a connotation of correctness that cannot be proven true. We disagree with this paradigm for a number of reasons. First, is the fact that whether the public agrees or not and whether the hydrogeologic community agrees or not, models are being used for regulatory decisions at a wide variety of sites, and many of these regulations call for some form of validation of the models. Therefore, instead of driving the process and studies to a halt, it is better to devote efforts to developing the tools and techniques that can be used for assessing the model results, and revising decisions based on them if needed. This would at least allow for allocating resources to achieve better understanding of the entire monitoring and validation process. Second, the term "model validation" requires as much effort to explain the underlying logic to the public as the terms "calibration," "history matching," and "benchmarking." To a technician or a mechanic who is familiar with calibrating digital scales or calipers, the term "calibration" alludes to high accuracy and correctness. Therefore, the calibration term can also be misleading to the public unless the underlying definitions and logic are clearly explained and simplified.

Third, Lee *et al.* (1996) identified significant misuses of groundwater models in 20 reviewed modeling reports that were used to make regulatory decisions. A well-established model validation procedure or process with trigger mechanisms for revisiting the model conceptualization if field data indicate deficiencies may have averted some of these misuses. Fourth, statements such as "groundwater models cannot be validated" may lead to a laid-back attitude on the part of modelers, hydrogeologists, or even regulatory agencies when it comes to testing and evaluating their models.

Finally, an analogy to the development and use of stochastic modeling can be made to support the above points. Dagan (2002) indicates that stochastic modeling of subsurface flow and transport has reached an advanced stage and has been applied to aquifer characterization, to the design and analysis of elaborate field experiments, and to a few major projects. So, the tools have been advanced and thoroughly developed despite the claims of many opponents who described the stochastic approaches as GIGO (garbage in garbage out). The lesson that can be learned here is that tools and techniques need to be developed and the focus needs to be shifted from what we call the process to how we best develop and shape the process of model validation in the hope that better decision making can be achieved.

REFERENCES

23 ERC 2091, Sixth Circuit. 1986. OHIO v. EPA, U.S. Court of Appeals, Sixth Circuit. 23 ERC 2091 – 23 ERC 2097.

Ababou, R., B. Sagar, and G. Wittmeyer. 1992. Testing procedures for spatially distributed flow models. *Advances in Water Resources* 15, 181-198.

Alley, W.M., and P.A. Emery. 1986. Groundwater model of the Blue River Basin, Nebraska – twenty years later. *Journal of Hydrology* 85, 225-250.

Anderson, M.G., and P.D. Bates. 2001. *Model Validation: Perspectives in Hydrological Science*. New York, NY: John Wiley & Sons, Ltd.

Anderson, M.P., and W.W. Woessner. 1992a. The role of postaudit in model validation. *Advances in Water Resources* 15, 167-173.

Anderson, M.P., and W.W. Woessner. 1992b. *Applied Ground Water Modeling: Simulation of Flow and Advective Transport*. New York, NY: Academic Press.

Andersson, K., B. Grundfelt, A. Larsson, and T. Nicolson. 1989. INTRAVAL as an integrated international effort for geosphere model validation – A status report. Proceedings of Symposium on Safety Assessment of Radioactive Waste Repositories, Paris, October 9-13, OECD Nuclear Energy Agency.

ASTM. 1993. Standard guide for comparing ground-water flow model simulations to site-specific information. Designation: D 5490 - 93 (Reapproved 2002), ASTM Internationsl, W. Conshohocken, PA.

AWR. 1992a. Special issue: Validation of Geo-Hydrological Models - Part 1. *Advances in Water Resources* 15, no. 1.

AWR. 1992b. Special issue: Validation of Geo-Hydrological Models - Part 2. *Advances in Water Resources* 15, no. 3.

Bair, S. 1994. Model (*in*)validation – A view from the courtroom. *Ground Water* 32, no. 4: 530-531.

Balci, O. 1988. Credibility assessment of simulation results: The state of the art, methodology and validation. *Simulation Series, The Society for Computer Simulation* 19, no. 1: 19-25.

Balci, O. 1989. How to assess the acceptability and credibility of simulation results. In *Proceedings of the 1989 Winter Simulation Conference*, Washington, DC. Edited by E. MacNair, K. Musselman, and P. Heidelberger.

Balci, O., and R.G. Sargent. 1981. A methodology for cost-risk analysis in the statistical validation of simulation models. *Communication of the ACM* 24, no. 4: 190-197.

Balci, O., and R.G. Sargent. 1982. Validation of multivariate response simulation models by using Hotelling's two-sample T^2 test. *Simulation* 39, no. 6: 185-192.

Balci, O., and R.G. Sargent. 1984. A bibliography on the credibility, assessment and validation of simulation and mathematical models. *Simuletter* 15, no. 3: 15-27.

Beljin, M.S. 1988. Testing and validation of models for solute transport in groundwater: Code intercomparison and evaluation of validation methodology. International Ground Water

Modeling Center, Holcomb Institute, Butler Univ., Indianapolis, Indiana. Report GWMI 88-11.

Boggs, J.M., S.C. Young, L.M. Beard, L.W. Gelhar, K.R. Rehfeldt, and E.E. Adams. 1992. Field study of dispersion in a heterogeneous aquifer, 1. Overview and site description. *Water Resources Research* 28, no. 12: 3281-3291.

Borgorinski, P., B. Blates, J. Larue, and K.H. Martens. 1988. The role of transport code verification and validation studies in licensing nuclear waste repositories in the FR of Germany. *Radiochimica Acta* 44-45, 367-372.

Bredehoeft, J.D., and L.F. Konikow. 1992. Reply to comment. *Advances in Water Resources* 15, 371-372.

Bredehoeft, J.D., and L.F. Konikow. 1993. Groundwater models: Validate or invalidate. *Ground Water* 31, no. 2: 178-179.

Brown, D.E., and A. Laase. 1995. Standard guide for calibrating a groundwater flow model application, Draft Section D18.21.10 Designation C-7, ASTM Standards on Ground Water and Vadose Zone Investigations, Ground Water Modeling.

Brown, D.M. 1996. Reducing modeling uncertainty using ASTM ground-water modeling standards. In *Subsurface Fluid-Flow* (*Ground-Water and Vadose Zone*) *Modeling, ASTM STP 1288,* eds. J.D. Ritchey and J.D. Rumbaugh, American Society for Testing and Materials, 24-41.

Broyd, T., D. Read, and B. Come. 1990. The CHEMVAL Project: An international study aimed at the verification and validation of equilibrium speciation and chemical transport computer programs. In *Proceedings of GEOVAL90 Symposium*, Stockholm, May 14-17, 1990, Swedish Nuclear Power Inspectorate (SKI), Stockholm.

Cacas, M.C., E. Ledoux, G. de Marsily, B. Tillie, A. Barbreau, E. Durand, B. Feuga, and P. Peaudecerf. 1990a. Modeling fracture flow with a stochastic discrete fracture network: Calibration and validation, 1. The flow model. *Water Resources Research* 26, no. 3: 479-489.

Cacas, M.C., E. Ledoux, G. de Marsily, A. Barbreau, P. Calmels, B. Gaillard, and R. Margrita. 1990b. Modeling fracture flow with a stochastic discrete fracture network: Calibration and validation, 2. The transport model. *Water Resources Research* 26, no. 3: 491-500.

Capilla J., J. Gómez-Hernández, and A. Sahuquillo. 1997. Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric data, 2. Demonstration on a synthetic aquifer. *Journal of Hydrology* 203, no. 1-4: 175-188.

Capilla J., J. Gómez-Hernández, and A. Sahuquillo. 1998. Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric data, 3. Application to the Culebra formation at the Waste Isolation Pilot Plant (WIPP), New Mexico, USA. *Journal of Hydrology* 207, no. 3-4: 254-269.

Chapman, J.B., K. Pohlmann, G. Pohll, A.E. Hassan, P. Sanders, M. Sanchez, and S. Jaunarajs. 2002. Remediation of the Faultless underground nuclear test: Moving forward in the face of model uncertainty. In *Proceedings of the Waste Management Conference, WM'02*, Tucson, Arizona.

Chapman, N., J. Anderson, P. Bogorinski, J. Carrera, J. Hadermann, D. Hodgkinson, P. Jackson, I. Neretnieks, S. Neuman, K. Skagius, T. Nicholson, C.F. Tsang, and C. Voss. 1994. Developing groundwater flow and transport models for radioactive waste disposal: Six years of experience from the INTRAVAL Project. In *GEOVAL 94, Validation Through Model Testing*, Proceedings of an NEA/SKI Symposium, Paris, France, 11-14 October, 45-58.

Cushman, J.H. 1997. *The Physics of Fluids in Hierarchical Porous Media: Angstroms to Miles*, Kluwer Acad., Norwell, Mass.

Dagan, G. 1989. *Flow and Transport in Porous Formation*s, Springer-Verlag, New York.

Dagan, G. 2002. An overview of stochastic modeling of groundwater flow and transport: From theory to applications. *EOS Transactions* 83, no. 53. American Geophysical Union.

D'Agnese, F.A., C.C. Faunt, M.C. Hill, and A.K. Turner. 1999. Death Valley regional groundwater flow model calibration using optimal parameter estimation methods and geoscientific information systems. *Advances in Water Resources* 22, no. 8: 777-790.

Davis, P.A., and M.T. Goodrich. 1990. A proposed strategy for the validation of groundwater flow and solute transport models. In *GEOVAL-90, Symposium on Validation of Geosphere Performance Assessment Models*, Stockholm, Sweden, 14-17 May, 580-588.

Davis, P.A., N.E. Olague, and M.T. Goodrich. 1991. Approaches for the validation of models used for performance assessment of high-level radioactive waste repositories. Sandia National Laboratories SAND90-0575, Albuquerque, New Mexico.

Davis, P.A., N.E. Olague, and M.T. Goodrich. 1992. Application of a validation strategy to Darcy's experiment. *Advances in Water Resources* 15, 175-180.

de Marsily, G. 1990. Validation of conceptual models of flow and transport in porous or fractured media. In *GEOVAL-90, Symposium on Validation of Geosphere Performance Assessment Models*, Stockholm, Sweden, 14-17 May, 36-50.

de Marsily, G., P. Combes, and P. Goblet. 1992. Comment on 'Groundwater models cannot be validated,' by L.F. Konikow and J. D. Bredehoeft. *Advances in Water Resources* 15, 367-369.

DOE. 1986. Environmental assessment – Yucca Mountain Site, Nevada Research and Development Area, Nevada. DOE/RW-0073, Vol. 2, U.S. Department of Energy, Office of Civilian Radioactive Waste Management, Washington, D.C.

Draper, N.R., and H. Smith. 1981. *Applied Regression Analysis, Second Edition*. John Wiley & Sons, New York.

Eisenberg, N., M. Federline, B. Sagar, G. Wittmeyer, J. Andersson, and S. Wingefors. 1994. Model validation from a regulatory perspective: A summary. In *GEOVAL 94, Validation Through Model Testing*, Proceedings of an NEA/SKI Symposium, Paris, France, 11-14 October, 421-434.

FFACO, 2000. Federal Facilities Agreement and Consent Order, Appendix VI: Corrective Action Strategy.

Flavelle, P. 1992. A quantitative measure of model validation and its potential use for regulatory purposes. *Advances in Water Resources* 15, 5-13.

Flavelle, P., S. Nguyen, and W. Napier. 1990. Lessons learned from model validation: A regulatory perspective. In *GEOVAL –1990: Symposium on Validation of Geosphere Flow and Transport Models*, Organization for Economic Cooperation and Development, Nuclear Energy Agency, Paris, France, 441-448.

Franks, S.W., and K.J. Beven. 1997. Bayesian estimation of uncertainty in land surface-atmosphere flux predictions. *Geophysical Research* 102, no. D20: 23991-23999.

Freer, J., K. Beven, and B. Ambroise. 1996. Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach. *Water Resources Research* 32, no. 7: 2161-2173.

Freyberg, D.L. 1986. A natural gradient experiment on solute transport in a sand aquifer, 2. Spatial moments and the advection and dispersion of nonreactive tracers. *Water Resources Research* 22, no. 13: 2031-2046.

Freyberg, D.L. 1988. An exercise in groundwater model calibration and prediction. *Ground Water* 26, no. 3: 350-360.

Frick, U. 1994. The Grimsel radionuclide migration experiment: A contribution to raising confidence in the validity of solute transport models used in performance assessment. In *GEOVAL 94, Validation Through Model Testing*, Proceedings of an NEA/SKI Symposium, Paris, France, 11-14 October, 245-272.

Gass, S.I. 1983. Decision-aiding models: Validation, assessment, and related issues for policy analysis. *Operations Research* 31, no. 4: 601-663.

Gass, S.I., and B.W. Thompson. 1980. Guidelines for model evaluation: An abridged version of the U.S. General Accounting Office exposure draft. *Operations Research* 28, no. 2: 431-479.

Gelhar, L.W. 1993. *Stochastic Subsurface Hydrology*, Prentice-Hall, Engle-wood Cliffs, New Jersey.

GEOVAL87. 1987. *Proceedings of Symposium on Verification and Validation of Geosphere Performance Assessment Models*, organized by Swedish Nuclear Power Inspectorate, Stockholm, Sweden, April 7-9.

GEOVAL90. 1990. *Proceedings of Symposium on Validation Geosphere Flow and Transport Models,* organized by Swedish Nuclear Power Inspectorate, Stockholm, Sweden, May 14-17.

GEOVAL94. 1994. *Proceedings of Symposium on Verification Through Model Testing*, organized by OECD Nuclear Energy Agency and the Swedish Nuclear Power Inspectorate, Paris, France, October 11-14.

Gómez-Hernández, J.J., A. Sahuquillo and J.E. Capilla. 1997. Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric data, 1. The theory. *Journal of Hydrology* 203, no. 1-4: 162-174.

Gorokhovski, V., and D. Nute. 1996. Validation of hydrogeological models is impossible: What's next? In *Calibration and Reliability in Groundwater Modeling*, K. Kovar and P. van der Hwijde (eds.), IAHS Publication no. 237: 417-424.

Grundfelt, B., B. Lindbom, A. Larsson, and K. Andersson. 1990. HYDROCOIN level 3 - Testing methods for sensitivity/uncertainty analysis. Proceedings of GEOVAL9O Symposium, Stockholm, May 14-17, 1990. Swedish Nuclear Power Inspectorate, (SKI), Stockholm.

Grundfelt, B. 1987. The HYDROCOIN Project - Overview and results from level one. Proceedings of International GEOVAL-87 Symposium, Swedish Nuclear Power Inspectorate, (SKI), Stockholm. April 7-9, 1987.

Hassan, A.E., K.F. Pohlmann, and J.B. Champan. 2001. Uncertainty analysis of radionuclide transport in a fractured coastal aquifer with geothermal effects. *Transport in Porous Media* 43, 107-136.

Hassan, A. E. 2002. Validation of numerical ground water models: A review. *Ground Water*, in review.

Hassanizadeh, S.M. 1990a. Experimental study of coupled flow and mass transport: A model validation exercise. In *ModelCARE 90: Calibration and Reliability in Groundwater Modeling*. IAHS Publ no. 195, 241-250.

Hassanizadeh, M. 1990b. Panel discussion. In *GEOVAL-90, Symposium on Validation of Geosphere Performance Assessment Models*, Stockholm, Sweden, 14-17 May, 631-658.

Herbert, A., W. Dershowitz, J. Long, and D. Hodgkinson. 1990. Validation of fracture flow models in the Stripa project. Proceedings of GEOVAL90 Symposium, Stockholm, May 14-17, Swedish Nuclear Power Inspectorate (SKI).

Hess, K.M., S.H. Wolf, and M.A. Celia. 1992. Large-scale natural gradient tracer test in sand and gravel, Cape Cod, Massachusetts, 3. Hydraulic conductivity variability and calculated macrodispersivities. *Water Resources Research* 28, no. 8: 2011-2027.

Hill, M.C., R.L. Cooley, and D.W. Pollock. 1998. A controlled experiment in groundwater flow model calibration. *Ground Water* 36, no. 3: 520-535.

Huyakorn, P.S., A.G. Kretschek, R.W. Broome, J.W. Mercer, and B.H. Lester. 1984. Testing and validation of models for simulating solute transport in groundwater. International Ground Water Modeling Center, Holcomb Research Institute, Butler University, Indianapolis, IN. Report GWMI 84-13.

IAEA. 1982. Radioactive waste management glossary. IAEA-TECDOC-264, International Atomic Energy Agency, Vienna, Australia.

IAEA, 1988. Radioactive waste management glossary, 2nd edition. IAEA-TECDOC-447, International Atomic Energy Agency, Vienna, Australia.

INTRACOIN. 1984. Final report level 1, Code verification. Report SKI 84:3, Swedish Nuclear Power Inspectorate, Stockholm, Sweden.

INTRACOIN. 1986. Final report levels 2 and 3, Model validation and uncertainty analysis. Report SKI 86:2, Swedish Nuclear Power Inspectorate, Stockholm, Sweden.

Jackson, C.P., D.A. Lever, and P.J. Summer. 1990. Validation of transport models for use in repository performance assessment: A view. In *GEOVAL-90, Symposium on Validation of Geosphere Performance Assessment Models*, Stockholm, Sweden, 14-17 May, 250-257.

Jackson, C.P., D.A. Lever, and P.J. Summer. 1992. Validation of transport models for use in repository performance assessment: A view illustrated for INTRAVAL test case 1b. *Advances in Water Resources* 15, 33-45.

Johnson, J.A. and D. J. Weimer. 1996. Verification of a ground water flow model application using recovery data and infiltration tests. In *Subsurface Fluid-Flow* (*Ground-Water and Vadose Zone*) *Modeling, ASTM STP 1288,* eds. J.D. Ritchey and J.D. Rumbaugh, American Society for Testing and Materials, 348-359.

Konikow, L.F. 1986. Prediction accuracy of a groundwater model: Lessons from a postaudit. *Ground Water* 24, no. 2: 173-184.

Konikow, L.F., and J.D. Bredehoeft. 1992. Groundwater models cannot be validated. *Advances in Water Resources* 15, 75-83.

Konikow, L.F., and J.D. Bredehoeft. 1993. The myth of validation in ground-water modeling. In *Proceedings 1993 Groundwater Modeling Conference*, IGWMC, Golden, CO, pp. A-4.

Kuhn, T.S. 1970. *The Structure of Scientific Revolution*. 2nd edition, University of Chicago Press, Chicago, Illinois.

Kuhn, T.S. 1982. Normal measurement and reasonable agreement. *Science in Context: Readings in the Sociology of Science*, MIT Press, 75-93.

LeBlanc, D.R., S.P. Garabedian, K.H. Hess, L.W. Gelhar, R.D. Quadri, K.G. Stollenwerk, and W.W. Wood. 1991. A large-scale natural gradient tracer test in sand and gravel, Cape Cod, Massachusetts, 1. Experimental design and observed tracer movement. *Water Resources Research* 27, no. 5: 895-910.

Lee, S.B., V. Ravi, J.R. Williams, and D.S. Burden. 1996. Evaluation of subsurface modeling application at CERCLA/RCRA sites. In *Subsurface Fluid-Flow* (*Ground-Water and Vadose Zone*) *Modeling, ASTM STP 1288,* eds. J.D. Ritchey and J.D. Rumbaugh, American Society for Testing and Materials, 3-13.

Legates, D.R., and G.J. McCabe Jr. 1999. Evaluating the use of goodness-of-fit measures in hydrologic and hydroclimatic model validation. *Water Resources Research* 35, no. 1: 233-241.

Leijnse, A., and S.M. Hassanizadeh. 1994. Model definition and model validation. *Advances in Water Resources* 17, 197-200.

Lewis, B.D., and F.S. Goldstein. 1982. Evaluation of a predictive groundwater solute transport model at the Idaho National Engineering Laboratory, Idaho. Water Resources Investigations Report 82-25, US Geological Survey, 71 pp.

Luis, S.J., and D. McLaughlin. 1992. A stochastic approach to model validation. *Advances in Water Resources* 15, 15-32.

Mackay, D.M., D.L. Freyberg, P.V. Roberts, and J.A. Cherry. 1986. A natural gradient experiment on solute transport in a sand aquifer, 1. Approach and overview of plume movement. *Water Resources Research* 22, no. 13: 2017-2029.

Maloszewski, P., and A. Zuber. 1992. On the calibration and validation of mathematical models for the interpretation of tracer experiments in groundwater. *Advances in Water Resources* 15, 47-62.

Maloszewski, P., and A. Zuber. 1993. Tracer experiments in fractured rocks: Matrix diffusion and the validity of models. *Water Resources Research* 29, no. 8: 2723-2735.

McCombie, C., I.G. McKinley, and P. Zuidema. 1990. Sufficient validation: The value of robustness in performance assessment and system design. In *GEOVAL-90, Symposium on Validation of Geosphere Performance Assessment Models*, Stockholm, Sweden, 14-17 May, 598-610.

McCombie, C., and I. McKinley. 1993. Validation – Another perspective. *Ground Water* 31, no. 4: 530-531.

McLaughlin, D., and S. Luis. 1990. A stochastic approach for validating models of unsaturated flow. In *GEOVAL-90, Symposium on Validation of Geosphere Performance Assessment Models*, Stockholm, Sweden, 14-17 May, 258-265.

Miller, I., and J.E. Freund. 1977. *Probability and Statistics for Engineers*. Prentice-Hall, New Jersey.

Miller, R.E., and P.K.M. van der Heijde. 1988. A groundwater research data center for model validation. Proceedings of the Indiana Water Resources Assoc. Symposium, June 8-10, Greencastle, IN.

Moltyaner, G.L., M.H. Klukas, C.A. Wills, and W.D. Killey. 1993. Numerical simulation of Twin Lake natural-gradient tracer tests: A comparison of methods. *Water Resources Research* 29, no. 10: 3433-3452.

Moran, M.S., and L.J. Mezgar. 1982. Evaluation factors for verification and validation of low-level waste disposal site models. Oak Ridge National Laboratory, Oak Ridge, TN. Report DOE/OR/21400-T119.

Mroczkowski, M., G.P. Raper, and G. Kuczera. 1997. The quest for more powerful validation of conceptual catchment models. *Water Resources Research* 33, no. 10: 2325-2335.

Mummert, M.C. 1996. Model validation and uncertainty analysis: An example using a nitrate percolation model. In *Subsurface Fluid-Flow* (*Ground-Water and Vadose Zone*) *Modeling, ASTM STP 1288,* eds. J.D. Ritchey and J.D. Rumbaugh, American Society for Testing and Materials, 187-200.

National Research Council (NRC). 2000. *Research Needs in Subsurface Science*. Washington, DC: National Academy Press.

Neuman, S.P. 1992. Validation of safety assessment models as a process of scientific and public confidence building. In *Proceedings of HLWM Conference*, Vol. 2, Las Vegas.

Nicholson, T.J. 1990. Recent accomplishments in the INTRAVAL Project - A status report on validation efforts. Proceedings of GEOVAL90 Symposium, Stockholm, Sweden, May 14-17, Swedish Nuclear Power Inspectorate (SKI).

Niederer, U. 1990. In search of truth: The regulatory necessity of validation. In *GEOVAL-90, Symposium on Validation of Geosphere Performance Assessment Models*, Stockholm, Sweden, 14-17 May, 29-35.

Oren, T. 1981. Concepts and criteria to assess acceptability of simulation studies. A Frame of Reference. *Communication of the ACM* 24, no. 4: 180-189.

Oreskes, N., K. Shrader-Frechette, and K. Belits. 1994. Verification, validation, and confirmation of numerical models in the earth sciences. *Science* 264, 641-646.

Person, M., and L.F. Konikow. 1986. Recalibration and predictive reliability of a solute-transport model of an irrigated stream-aquifer system. *Journal of Hydrology* 87, 145-165.

Pescatore, C. 1994. Validation: An overview of definitions. In *GEOVAL 94, Validation Through Model Testing*, Proceedings of an NEA/SKI Symposium, Paris, France, 11-14 October, 15-24.

Poeter, E.P., and M.C. Hill. 1997. Inverse models: A necessary next step in groundwater modeling. *Ground Water* 35, no. 2: 250-260.

Pohll, G., A.E. Hassan, J.B. Chapman, C. Papelis, and R. Andricevic. 1999. Modeling groundwater flow and radioactive transport in a fractured aquifer. *Ground Water* 37, no. 5: 770-784.

Pohll, G., and T. Mihevc. 2000. Data Decision Analysis: Central Nevada Test Area. Publication No. 45179. Reno, Nevada: Desert Research Institute, Division of Hydrologic Sciences.

Pohlmann, K.F., A.E. Hassan, and J.B. Chapman. 1999. Evaluation of groundwater flow and transport at the Faultless underground nuclear test, Central Nevada Testing Area. Publication No. 45165. Las Vegas, Nevada: Desert Research Institute, Division of Hydrologic Sciences.

Pohlmann, K.F., A.E. Hassan, and J.B. Chapman. 2000. Description of hydrogeologic heterogeneity and evaluation of radionuclide transport at an underground nuclear test. *Contaminant Hydrology* 44, 353-386.

Pohlmann, K.F., G. Pohll, J. Chapman, A. Hassan, R. Carroll, and C. Shirley. 2001. Modeling of groundwater contaminant boundaries for the Shoal underground nuclear test. Publication No. 45184. Las Vegas, Nevada: Desert Research Institute, Division of Hydrologic Sciences.

Popper, K.R. 1968. *The Logic of Scientific Discovery*, Harper and Row Publishers Inc., New York.

Raven, K.G., K.S. Novakowski, and P.A. Lapcevic. 1988. Interpretation of field tracer tests of a single fracture using a transient solute storage model. *Water Resources Research* 24, no. 12: 2019-2032.

Robertson, J.B. 1974. Digital modeling of radioactive and chemical waste transport in the Snake River Plain aquifer at the National Reactor Testing Station, Idaho. Open File Report IDO-22054, US Geological Survey.

Rogers, P. 1978. On the choice of the 'appropriate model' for water resources planning and management. *Water Resources Research* 14, no. 6: 1003-1010.

Sahuquillo, A., J. Capilla, J.J. Gómez-Hernández and J. Andreu. 1992. Conditional simulation of transmissivity fields honouring piezometric data. In Blair, W. R. and Cabrera, E. (Eds), *Hydraulic Engineering Software IV, "Fluid Flow Modeling,"* Computational Mech. Publ., Boston, and Elsevier Applied Science, London.

Sargent, R.G. 1984. Simulation Model Validation. In *Simulation and Model-based Methodologies: An Integrative View*. Edited by Oren *et al.,* Springer-Verlag.

Sargent, R.G. 1988. A tutorial on validation and verification of simulation models. Proceedings of 1988 Winter Simulation Conference. Edited by M. Abrams, P. Haigh, and J. Comfort, 33-39.

Sargent, R.G. 1990. Validation of mathematical models. In *GEOVAL-90, Symposium on Validation of Geosphere Performance Assessment Models*, Stockholm, Sweden, 14-17 May, 571-579.

Schlesinger, S. 1979. Terminology for model credibility. *Simulation* 32, no. 3: 103-104.

Schruben, L.W. 1980. *Establishing the Credibility of Simulations - The Art and the Science*. Prentice-Hall.

Shah Alam, A.H.M. 1998. Regulatory guidance for accepting contaminant fate and transport models. In *Proceedings Modflow 98*, ed. E. Poeter et al. 387-393. Goldon, CO.

Shapiro, A.M., and J.R. Nicholas. 1989. Assessing the validity of the channel model of fracture aperture under field conditions. *Water Resources Research* 25, no. 5: 817-828.

SSI (Swedish National Institute of Radiation Protection). 1990. *Proceedings of Symposium and Workshop on the Validity of Environmental Transfer Models (BIOMOVS)*, October 8-12, 1990. Swedish National Institute of Radiation Protection (SSI), Stockholm.

Swedish Nuclear Power Inspectorate. 1987. The International HYDROCOIN Project—Background and Results. OECD, Paris.

Swedish Nuclear Power Inspectorate. 1990. The International INTRAVAL Project—Background and Results. OECD, Paris.

Tsang, C.F. 1987. Comments on model validation. *Transport in Porous Media* 2, no. 6: 623-630.

Tsang, C.F. 1989. Tracer travel time and model validation. *Radioactive Waste Management and the Nuclear Fuel Cycle* 13, no. 1-4: 311-323.

Tsang, C.F. 1991. The modeling process and model validation. *Ground Water* 29(6), 825-831.

Tsang, C.F. 1994. Validation and technical issues from GEOVAL-90 to GEOVAL-94. In *GEOVAL 94, Validation Through Model Testing*, Proceedings of an NEA/SKI Symposium, Paris, France, 11-14 October, 27-33.

U.S. Nuclear Regulatory Commission (USNRC). 1984. A revised modeling strategy document for high-level waste performance assessment. U.S. Nuclear Regulatory Commission, Washington, D.C.

van der Heijde, P.K.M., P.S. Huyakorn, and J.W. Mercer. 1985. Testing and validation of groundwater models. Proceedings of the NWWA/IGWMC Conference on "Practical Applications of Ground Water Models." August 19-20, Columbus, OH.

van der Heijde, P.K.M. 1987. Quality assurance in computer simulations of groundwater contamination. *Environmental Software* 2, no. 1.

van der Heijde, P.K.M., W.I.M. Elderhorst, R.A. Miller, and M.F. Trehan. 1989. The establishment of a groundwater research data center for validation of subsurface flow and transport models. International Ground Water Modeling Center, Holcomb Research Institute, Butler Univ., Indianapolis, IN. Report GWMI 89-01.

van der Heijde, P.K.M. 1990. Quality assurance in the development and application of ground water models. In *ModelCARE 90: Calibration and Reliability in Groundwater Modeling* IAHS Publ. no. 195, 271-278.

van der Heijde, P.K.M., and D.A. Kanzer. 1997. Groundwater model testing: Systematic evaluation and testing of code functionality and performance. US Environmental Protection Agency, National Risk Management Research Laboratory, Ada, Oklahoma. Report EPA/600/R-97/007. http://www.epa.gov/ada/download/project/gwtestng.pdf (Accessed March 2002).

Voss, C.F. 1990. A proposed methodology for validating performance assessment models for the DOE Office of Civilian Radioactive Waste Management Program. In *High Level Radioactive Waste Management*, Vol. 1. American Society of Civil Engineers, New York, 359-363.

Weaver, J. D., R.K. Digel, and P.V. Rosasco. 1996. A postaudit of ground water flow models used in design of a ground water capture/containment system. In *Subsurface Fluid-Flow* (*Ground-Water and Vadose Zone*) *Modeling, ASTM STP 1288,* eds. J.D. Ritchey and J.D. Rumbaugb, American Society for Testing and Materials, 377-390.

Wen X.-H., J. Gómez-Hernández, J.E. Capilla, and A. Sahuquillo. 1996. Significance of conditioning to piezometric head data for predictions of mass transport in groundwater modeling. *Mathematical Geology* 28, no. 7: 951-968.

Wen X.-H., J.E. Capilla, C.V. Deutsch, J. Gómez-Hernández, and A.S. Cullick. 1999. A program to create permeability fields that honor single-phase flow rate and pressure data. *Computer & Geosciences* 25, 217-230.

Willmott, C.J. 1981. On the validation of models. *Physical Geography* 2, 184-194.

Willmott, C.J., S.G. Ackleson, R.E. Davis, J.J. Feddema, K.M. Klink, D.R. Legates, J. O'Donnell, and C.M. Rowe. 1985. Statistics for the evaluation and comparison of models. *Journal of Geophysical Research* 90, 8995-9005.

Woessner, W.W. and M. P. Anderson. 1996. Good model-bad model, understanding the flow modeling process. In *Subsurface Fluid-Flow* (*Ground-Water and Vadose Zone*) *Modeling, ASTM STP 1288,* eds. J.D. Ritchey and J.D. Rumbaugh, American Society for Testing and Materials, 14-23.

Young, P. 2001. Data-based mechanistic modeling and validation of rainfall-flow processes. In *Model Validation: Perspectives in Hydrological Science*, eds. M.G. Anderson and P.D. Bates, Chichester: J. Wiley, 117-161.

Zeigler, B.P. 1976. *Theory of Modeling and Simulation*. John Wiley and Sons, Inc., New York.

Zhang, D. 1998. Numerical solutions to statistical moment equations of groundwater flow in nonstationary, bounded heterogeneous media. *Water Resources Research* 34, no. 3: 529-538.

Zimmerman, D.A., G. de Marsily, C.A. Gotway, M.G. Marietta, C.L. Axness, R.L. Bras, J. Carrera, G. Dagan, P.B. Davies, D.P. Gallegos, A. Galli, J. Gómez-Hernández, P. Grindgrod, A.L. Gutjahr, P.K. Kitanidis, A.M. LaVenue, D. McLaughlin, S.P. Neuman, B.S. RamaRao, C. Ravenne and Y. Rubin. 1998. A comparison of seven geostatistically based inverse approaches to estimate transmissivities for modeling advective transport by groundwater flow. *Water Resources Research* 34, no. 6: 1373-1413.

Zuidema, P. 1994. Validation: Demonstration of disposal safety requires a practicable approach. In *GEOVAL 94, Validation Through Model Testing*, Proceedings of an NEA/SKI Symposium, Paris, France, 11-14 October, 35-42.

Appendices: Background and Theoretical Concepts

Appendix A: Generalized Likelihood Uncertainty Estimate Analysis

To honor site-specific data during calibration and subsequent modeling, the generalized likelihood uncertainty estimator (GLUE) algorithm can be used (Freer *et al.*, 1996; Franks and Beven, 1997; Pohlmann *et al.*, 2001). The GLUE procedure is an extension of Monte Carlo random sampling to incorporate the goodness-of-fit of each simulation. A likelihood measure is an evaluation of the quantitative goodness-of-fit. For example, the likelihood estimator for the solution of the flow equations can be defined as

$$L(Y|\Theta) = \left[\sum (\varepsilon)^2\right]^{-M} \tag{A1}$$

where

$$\varepsilon = h_j^* - \hat{h}_j \tag{A2}$$

and $L(Y|\Theta)$ is the likelihood of the vector of outputs, Y, knowing Θ, the vector of random inputs; $\hat{h}_j$ is the simulated head at the point j; h_j^* is the observed head at that point; and M is a likelihood shape factor. Although the choice of the M factor is subjective, its value defines its relative function. As M approaches zero, likelihood approaches unity and each simulation has equal weight, as is the case with traditional Monte Carlo analysis. As M approaches infinity, simulations with the lowest sum of squared errors (the simulations that best fit the field data) receive essentially all of the weight, which is analogous to an inverse solution. The likelihood weights that are calculated for each realization based on Eq. (A1) can be used in subsequent modeling to give more weight to those realizations that best fit the field data during the calibration process. Also, these weights can be used later in the validation stage to compare the performance of individual realizations when acquiring new field data for validation analysis.

As was shown in Figure 3, one of the steps in the proposed validation approach is to quantitatively evaluate the calibration goodness-of-fit for each realization using the GLUE analysis. This analysis will give each realization a relative weight indicating its strength in matching the calibration targets. These weights are the first quantitative measures for different realizations, which can later be combined with and compared to the different evaluations and tests performed using the validation data.

Appendix B: Goodness-of-Fit Measures/

Legates and McCabe (1999) provide an evaluation of the common goodness-of-fit measures that are used in hydrologic and hydroclimatic model validation. They argue that correlation and correlation-based measures (e.g., coefficient of determination R^2) are oversensitive to extreme values or outliers and insensitive to additive and proportional differences between model predictions and observations. They conclude that these measures should not be used to assess goodness-of-fit of a hydrologic model and that additional evaluations such as summary statistics and absolute error measures should supplement model evaluation tools. They also present useful alternative goodness-of-fit and relative error measures (e.g., coefficient of efficiency, index of agreement) that overcome many of the limitations of correlation-based measures. The remainder of this Appendix is a summary of the presentation of Legates and McCabe (1999) highlighting the definitions and differences between different measures as presented in the context of model evaluation.

B.1 *Coefficient of Determination R^2*

The coefficient of determination describes the proportion of the total variance in the observed data that can be explained by the model and ranges from 0.0 to 1.0, with higher values indicating better agreement

$$R^2 = \frac{\sum_{i=1}^{N}(O_i - \overline{O})(P_i - \overline{P})}{\left[\sum_{i=1}^{N}(O_i - \overline{O})^2\right]^{0.5}\left[\sum_{i=1}^{N}(P_i - \overline{P})^2\right]^{0.5}} \qquad \text{(B1)}$$

where the overbar denotes the mean, P denotes predicted variable, O indicates observed values and N is the number of available pairs of predicted versus measured values. It can be seen that if $P_i = (AO_i + B)$ for any non-zero value of A and any value of B, then $R^2 = 1.0$. Thus R^2 is insensitive to additive and proportional differences between the model predictions and observations. It is also more sensitive to outliers than to observations near the mean.

B.2 *Coefficient of Efficiency E*

The coefficient of efficiency, which ranges from minus infinity to 1.0, is defined as (Legates and McCabe, 1999)

$$E = 1 - \frac{\sum_{i=1}^{N}(O_i - P_i)^2}{\sum_{i=1}^{N}\left(O_i - \overline{O}\right)^2} \qquad \text{(B2)}$$

The coefficient of efficiency represents an improvement over the coefficient of determination for model evaluation purposes in that it is sensitive to differences in the observed and model-simulated means and variances; that is, if $P_i = (AO_i + B)$, then E decreases as A and B vary from 1.0 and 0.0, respectively. Because of the squared differences, however, E is overly sensitive to extreme values, as is R^2.

B.3 ***Index of Agreement d***

The index of agreement, *d*, was developed to overcome the insensitivity of correlation-based measures to additive and proportional differences between observations and model simulations. It is expressed as (Willmott, 1981)

$$d = 1 - \frac{\sum_{i=1}^{N}(O_i - P_i)^2}{\sum_{i=1}^{N}\left(\left|P_i - \overline{O}\right| + \left|O_i - \overline{O}\right|\right)^2} = 1 - N\,\frac{\text{MSE}}{\text{PE}} \qquad \text{(B3)}$$

The index of agreement varies from 0.0 to 1.0 and represents the ratio between the mean square error and the "potential error" (PE), multiplied by *N* and then subtracted from unity. The potential error represents the largest value that $(O_i - P_i)^2$ can attain for each observed-simulated pair (Legates and McCabe, 1999). As with *E*, the index of agreement, *d*, represents an improvement over R^2, but also is sensitive to extreme values owing to the squared differences.

The sensitivity of R^2, *E* and *d* to extreme values led to the suggestion that a more generic index of agreement could be used in the form (Willmott *et al.*, 1985)

$$d_j = 1 - \frac{\sum_{i=1}^{N}(O_i - P_i)^j}{\sum_{i=1}^{N}\left(\left|P_i - \overline{O}\right| + \left|O_i - \overline{O}\right|\right)^j} \qquad \text{(B4)}$$

where *j* represents an arbitrary power (i.e., a positive integer). The original index of agreement *d* given in Eq. (B3) becomes d_2 using this notation. For $j = 1$, the resulting index, d_1, has the advantage that errors and differences are given their appropriate weighting, not inflated by their squared values. Similarly, the coefficient of efficiency can be adjusted as

$$E_j = 1 - \frac{\sum_{i=1}^{N}(O_i - P_i)^j}{\sum_{i=1}^{N}\left(O_i - \overline{O}\right)^j} \qquad \text{(B5)}$$

Legates and McCabe (1999) and Willmott (1981) argue that these dimensionless measures (e.g., E_1 and d_1) should not be used exclusively. It may be necessary and appropriate to quantify the error in terms of the units of the variable at hand. Therefore, in addition to *E* and *d* measures, one has to consider absolute error measures, which include the root mean square error ($\text{RMSE} = \sqrt{\text{MSE}}$) and the mean absolute error $\left(\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|O_i - P_i\right|\right)$. These additional measures describe the difference between the model simulations and observations in the units of the variable predicted. Legates and McCabe (1999) conclude by recommending that the assessment of the model performance should include at least one "goodness-of-fit" or relative error measure (e.g., E_1 and d_1) and at least one absolute error measure (e.g., RMSE or MAE)

with additional supporting information (e.g., a comparison between the observed and simulated mean and standard deviations).

It should be mentioned here that the analysis of Legates and McCabe (1999) was based on analysis of time series models, for which large-size data sets are available to test prediction models. In the subsurface, however, availability of such abundant data never is (and never will be) the case. Therefore, some of the goodness-of-fit measures discussed above may not be usable for such limited data. It is thus important not to rely on a single measure, but to use as many measures as possible to get a better evaluation of the model predictions.

Appendix C: Linear Regression Analysis

Davis and Goodrich (1990) propose examining the deviations from model calibration for trends to identify systematic errors, the existence of which would invalidate the model. Flavelle *et al.* (1990) perform a linear regression analysis of calculated versus measured data for both the calibration and the validation processes. They interpret the standard error of the regression as the goodness-of-fit and the slope of the regression line as the model bias. Flavelle (1992) argues that this linear regression analysis and its interpretation are the initial steps for evaluating model validation. From the perspective of making regulatory decisions based on model calculations, this approach has some advantages and is based on the following reasoning summarized from Flavelle (1992).

Three components can be looked at, which lead to four possibilities for the linear regression analysis and interpretation. The three components are the input data used in the model, the model itself (all of the necessary processes, mechanisms and structures) and the validation or calibration data. The first possibility is that the three components are perfect, i.e., perfectly known input data (no uncertainty) are applied to a perfect model and the calculated results are compared to perfect observations (no uncertainty and no randomness). In this case, a plot of calculated versus measured data would be a perfect straight line with unit slope, zero intercept, perfect correlation coefficient and no regression error (Figure C1-a, recreated based on Figure 1 of Flavelle (1992)). The second possibility occurs if the model is not perfect, whereas both input data and validation/calibration data are perfect. A systematic (i.e., non-random) bias would occur and the regression line would have a slope different from unity and/or an intercept different from zero (Figure C1-b recreated based on Figure 1 of Flavelle (1992)). The data points may not be collinear, so the correlation coefficient may be less than one. The third possibility occurs when the input data are uncertain, and/or when the observations are uncertain or have a random component. In this case, the results from a perfect model would have a regression line with a unit slope but with some data scatter about the line (Figure C1-c). This data scatter is measured by the standard error of the regression, which is used to determine the confidence interval about the regression line. Finally, if the observed or input data are uncertain and the model is not perfect, the regression line will not have a unit slope and/or intercept of zero, and there will be data scatter about the regression line (Figure C1-d).

Following this reasoning, a linear regression of calculated against measured data provides an initial method to evaluate empirically the quality of the data fit. Bias in the model and uncertainty in the input and measured data would be expected to affect both the slope of the regression line and the standard error of the regression. There are several techniques for fitting a straight line through x, y data using regression analysis. The most common regression analysis for predictive purposes (and the most common regression analysis in general) is the Ordinary Least Squares (OLS) regression of a dependent variable against an independent variable.

Based on this linear regression, one needs to statistically test the assertion that the slope of the regression line is unity and that the intercept of the line is zero. Hypothesis testing (see Appendix D) can be used for this purpose with the null hypothesis for the slope being H_0**:** slope = 1, and the alternate hypothesis is H_1**:** slope ≠ 1. The test statistic is ((slope-1) ÷ standard deviation of the slope). This is to be compared to the critical value of the t-distribution at (n - 2) degrees of freedom (n is the number of data points) and at the α level of significance, $t(n-2,1-0\cdot 5\alpha)$. If the absolute value of the test statistic exceeds the critical value, the null

hypothesis is rejected. In a similar manner, the null hypothesis of a zero intercept can be examined. Failing to reject both null hypotheses does not mean the model is free of biases, only that this analysis fails to identify any bias (Flavelle, 1992).

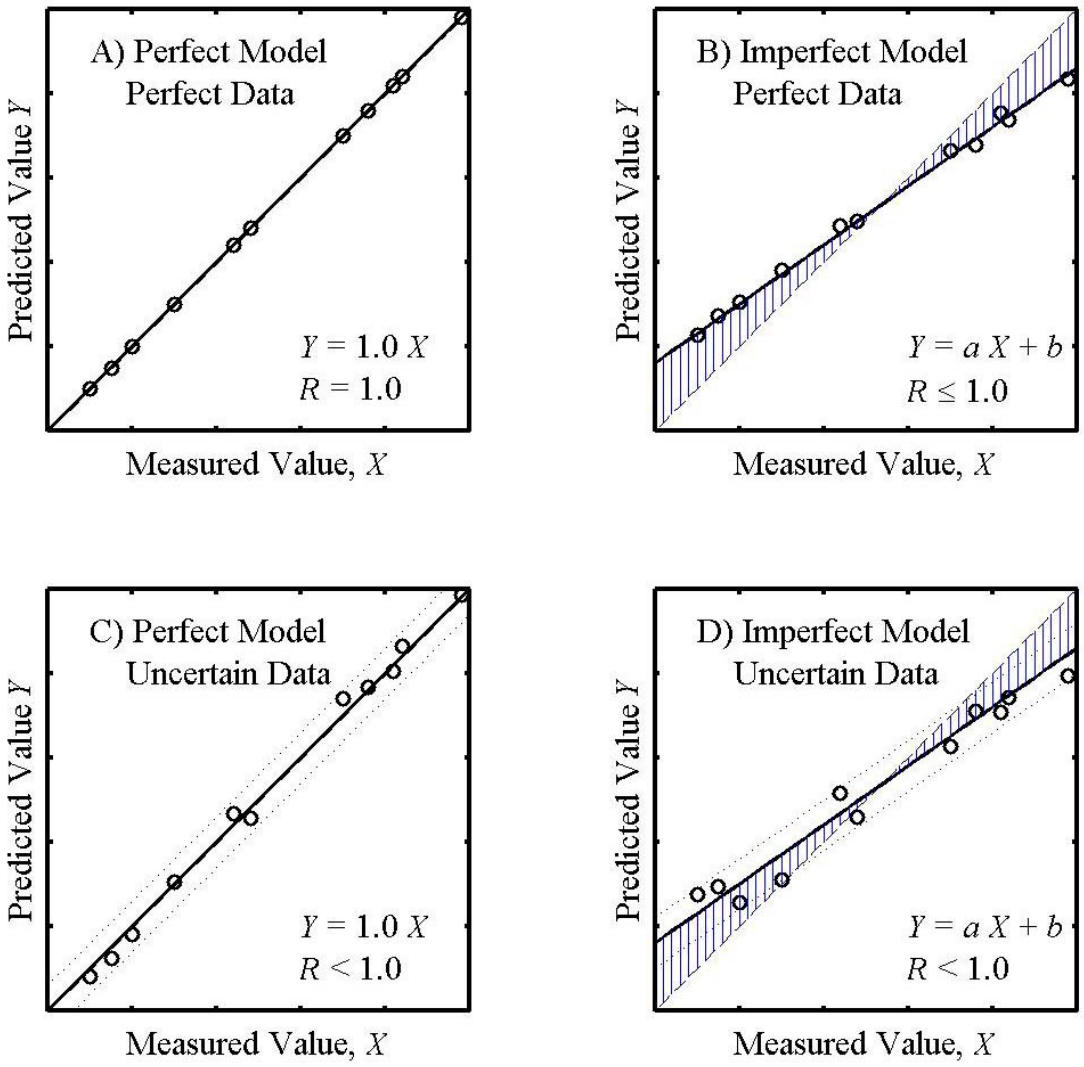

Figure C1. Linear regression scenarios when applied to the comparison between model predictions and observations (adapted from Flavelle, 1992).

The use of linear regression analysis to evaluate the accuracy of model calculations is not inconsistent with some of the other calibration approaches being developed. Most of these other approaches use a goodness-of-fit parameter based on some form of the difference between the calculated and measured values. Davis and Goodrich (1990) suggested that the deviations of the calculated values from the observations should be examined for trends to identify model bias. The deviations between calculated and observed values correspond to the deviation of observed versus predicted data points from the 45° line on the linear plot. Trends in the set of deviations are what cause the slope of a regression line to vary from unity. Regression analysis has a compelling advantage over analysis of the deviations, as it has been shown that the assumption

that the regression residuals are normally distributed is not unreasonable (Draper and Smith, 1981), while the deviations between calculated and observed data may not be normally distributed. Statistical analysis of non-normally distributed data usually requires non-parametric statistical tests, which are more complex than parametric tests used for normally distributed data (Flavelle, 1992).

Appendix D: Hypothesis Testing

Statistical hypothesis testing can be used as a quantitative tool for evaluating predictive models. The test usually postulates a null hypothesis (H_0) and a complementary hypothesis (H_1). The null hypothesis postulates the assumption or result that needs to be tested (e.g., the model is valid or the linear regression line has a slope of unity), while the complementary hypothesis postulates the opposite. Two types of errors can occur in hypothesis testing with certain probabilities: type I error and type II error. The probability of type I error is called model builder's risk (α), whereas the probability of type II error is called model user's risk (β), and in model validation, model user's risk is extremely important and must be kept small (Sargent, 1990). These probabilities and those for making the right decisions are shown in Table (D1), adapted from Balci and Sargent (1981). Both type I and type II errors must be considered in using hypothesis testing for model validation and the risks resulting from these errors can be decreased at the expense of increasing the sample sizes of observations.

Table D1. Outcomes of hypothesis testing (adapted from Balci and Sargent, 1981).

Result of Hypothesis Testing	Actual Status of the Model	
	Model is Valid Null Hypothesis, H_0, is True	Model is Invalid Complementary Hypothesis H_1 is True
Do not reject H_0	Correct Decision	Model User's Risk β
Reject H_0	Model Builder's Risk α	Correct Decision

Balci and Sargent (1981, 1982) developed a methodology for constructing the relationships between model user's risk, model builder's risk, acceptable validity range, sample sizes and cost of data collection when statistical hypothesis testing is used for validating a simulation model of a real, observable system. The acceptable validity range is the amount of acceptable accuracy required for the model to be valid under a given set of experimental conditions. This range is determined in terms of a validity measure that determines the amount of agreement (or lack thereof) between the model predictions and the actual observable system. Balci and Sargent (1981) use an Operating Characteristic Curve (Miller and Freund, 1977) to examine the probability of accepting a simulation model as being valid and the interplay between validity measure, model builder's risk and model user's risk. This Operation Characteristic Curve is shown in Figure D1 for two different values of confidence level, α^*. It can be easily seen from the figure that the model builder's risk has the limits $\alpha^* \le \alpha \le (1-\beta^*)$ and the model user's risk has the limits $0 \le \beta \le \beta^*$. Decreasing the upper bound of the model user's risk increases the upper bound of the model builder's risk. One can decrease model user's risk by increasing the range of acceptable validity (increasing λ^*), increasing the minimum model builder's risk α^*, or increasing the sample size of the observation data. Thus, there is a direct relation between model builder's risk, model user's risk, acceptable validity range and sample

sizes of observations (equivalent to a cost parameter), and a tradeoff among these parameters can be made by the model sponsor, model user (or regulator), and model builder for the intended application of the model (Balci and Sargent, 1981).

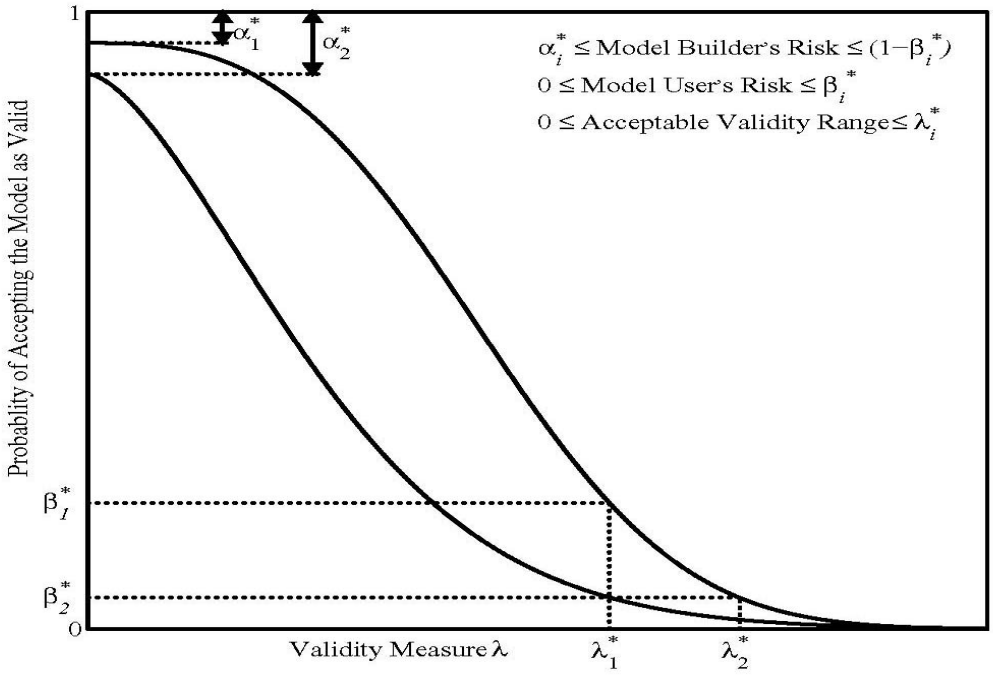

Figure D1. Schematic representation of the Operating Characteristic Curves depicting the relationships between α^*, β^*, and γ^* (adapted from Balci and Sargent, 1981).

Appendix E: Stochastic Validation Approach (Luis and McLaughlin, 1992)

Luis and McLaughlin (1992) propose and apply a stochastic approach that relies on hypothesis testing to validate a two-dimensional, deterministic, unsaturated flow model for predicting moisture movement at a field site near Las Cruces, New Mexico. The approach begins by identifying the factors that contribute to the differences between model predictions and observations. A number of assumptions are used in Luis and McLaughlin's (1992) study and are invoked here and adapted for the general case of a saturated flow model and the special case of the CNTA flow model. It is assumed that a flow model is used for predicting the distribution of hydraulic head in space, which describes the large-scale flow behavior that affects the movement and transport of contaminants. Another assumption is that the observations to be made for the purpose of model validation are small-scale observations collected at sparse points in space and are assumed to be consistent with the steady state assumption of the model.

Under these assumptions, the differences between predicted and measured head values can be attributed to the following three error sources: (1) measurement errors which represent the difference between the true values and the small-scale values of hydraulic head; (2) spatial heterogeneity, which represents the difference between the large-scale trend (or smoothed head) that the model is intended to predict and the true small-scale values of head; and (3) model error, which represents the difference between the model prediction and the actual smoothed trend. Figure E1-A shows schematic representations of these error sources, where an actual, h_j, fluctuating (due to heterogeneity) head distribution with a large-scale trend, $\bar{h}_j$, is shown in conjunction with a hypothesized stepwise distribution representing model prediction, $\hat{h}_j$. Measurement errors are only dependent on the measurement protocol and accuracy of the device used, which are not related in any way to the model. Spatial heterogeneity effect is embedded in the difference between the small-scale measurements and the large-scale trend, and this difference is not really an error but a reflection of the difference in scale between the measured and predicted quantities (McLaughlin and Luis, 1990). Model error is a reflection of the model's ability to predict the large-scale trend, which is the primary quantity of interest in this case, and could be due to conceptual deficiencies or erroneous inputs.

The first step in the analysis will be to decompose residuals into three terms, which account for the three error sources identified earlier. The j^{th} measurement residual, ε_j, observed at location $\mathbf{x}_j$ (for j= 1,...N), where N is the total number of head measurements used for validation, can be written as

$$\varepsilon(\mathbf{x}_j) = \varepsilon_j = h_j^* - \hat{h}_j(\hat{\eta}) \qquad \text{(E1)}$$

where $h_j^* = h^*(\mathbf{x}_j)$ is the head measurement at $\mathbf{x}_j$ and $\hat{h}_j = \hat{h}(\mathbf{x}_j | \hat{\eta})$ is the model prediction at the same location obtained by using a set of estimated model parameters, $\hat{\eta}$. Note that the hat symbol is used to refer to estimated or model-predicted quantities. Equation (E1), representing the mismatch between observations and model predictions, can be rewritten in terms of three components of the error or the mismatch. This leads to the equation

$$\varepsilon_j = [h_j^* - h_j] + [h_j - \bar{h}_j] + [\bar{h}_j - \hat{h}_j(\hat{\eta})] \qquad \text{(E2)}$$

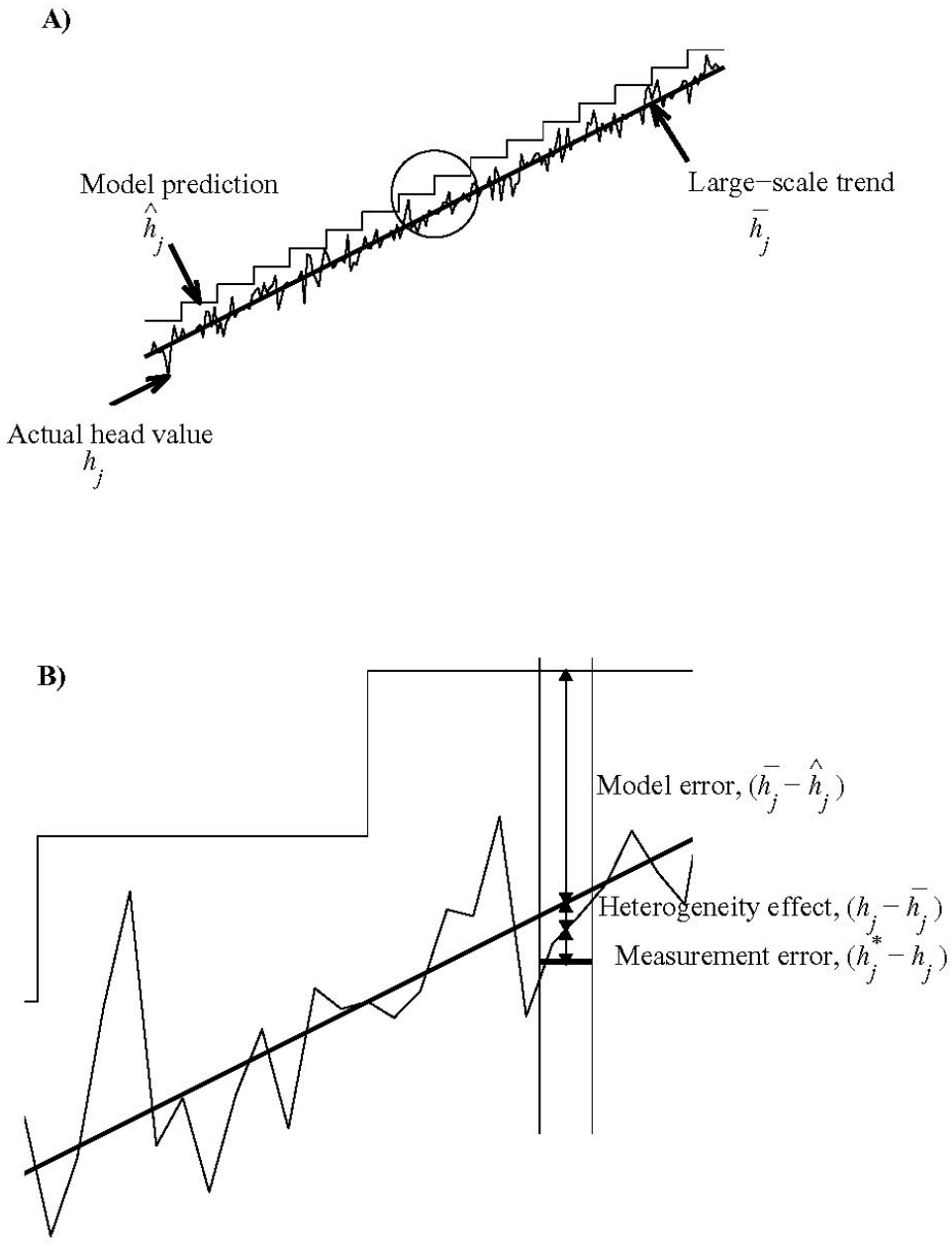

Figure E1. Schematic representations of the actual head distribution, large-scale trend, and stepwise model prediction (A), and the decomposition of the measurement residual into three error sources or components (B).

where the first term between the square brackets represents measurement error, the second bracketed term represents the effect of geologic heterogeneity, and the last term represents the model error. In (E2), $h_j = h(\mathbf{x}_j)$ is the true head value at $\mathbf{x}_j$ and $\overline{h}_j = \overline{h}(\mathbf{x}_j)$ is the smoothed value of the large-scale trend or the expected value of h_j. Again it is assumed here that the mathematical expectation of the head represents the large-scale (e.g., at the 50-m-grid scale of the CNTA model) values of head that govern the flow pattern and the transport velocities. Equation (E2) now defines the separate errors contributing to the differences between measurements and predictions. These errors are schematically shown in Figure E1-B.

The second step is to consider the hypothesis that the model prediction is equal to the smoothed, large-scale values. This is equivalent to assuming that the model error term in (E2) is zero. In statistical terms the following null hypothesis is considered

$$\begin{aligned} &H_0 : \text{Model error is negligible, } \hat{h}_j(\hat{\eta}) = \overline{h}_j \\ &H_1 : \text{Model error is significant, } \hat{h}_j(\hat{\eta}) \neq \overline{h}_j \end{aligned} \qquad \text{(E3)}$$

To apply this hypothesis testing technique to the model validation problem, one must find test statistics that can be used to check the hypothesis defined in (E3). These statistics should depend on available head measurements and should be designed to minimize the risk associated with making erroneous decisions of hypothesis testing (see Appendix D on hypothesis testing and associated errors). If one designs a very stringent test, the model user's risk, β, will be small and the model builder's risk, α, will be large (i.e., it will tend incorrectly to reject good models). If, on the other hand, the test is less stringent, it will have a large β and a small α (i.e., it will tend incorrectly to accept bad models).

Luis and McLaughlin (1992) point out that there is no rigorous way to develop an optimal test for the spatially distributed hypothesis-testing problem posed above. A number of reasonable tests, which can capture the different aspects of model evaluation and its inadequacy, can be used instead. Luis and McLaughlin propose a quantitative approach to determine whether statistics such as the sample mean and covariance of the residuals are consistent with hypothesis H_0 in (E3). When the hypothesis is true, the mean measurement residual is written as

$$\overline{\varepsilon}_j = \overline{[h_j^* - h_j]} + \overline{[h_j - \overline{h}_j]} \qquad \text{(E4)}$$

The second term on the right-hand side is zero by construction. The first term is simply the mean measurement error (the measurement bias). If the bias in the measurement device is known, then it can replace the first term on the right-hand side of (E4). Otherwise, it is expected that the measurement residuals have a zero mean at all sample locations

$$\overline{\varepsilon}_j = 0 \qquad \forall\ j \qquad \text{(E5)}$$

To derive the variance of the measurement residual, Luis and McLaughlin (1992) assume that measurement errors $(h_j^* - h_j)$ are uncorrelated with errors due to spatial heterogeneity $(h_j - \overline{h}_j)$ when the mean measurement residual $(\overline{\varepsilon}_j)$ is zero. The covariance between two different measurement residuals is then written as

$$P_{\varepsilon\varepsilon}(j,k) = \overline{[h_j^* - h_j][h_k^* - h_k]} + \overline{[h_j - \overline{h}_j][h_k - \overline{h}_k]} \qquad \text{(E6)}$$

If it is further assumed that measurement errors at different locations are uncorrelated with one another and have a common variance, Eq. (E6) can be written as

$$P_{\varepsilon\varepsilon}(j,k) = \sigma_{h^*}^2 \delta_{jk} + P_{hh}(j,k) \tag{E7}$$

where $\sigma_{h^*}^2$ is the measurement error variance, $P_{hh}(j, k)$ is the covariance between the true point head measurements h_j and h_k , and δ_{jk} is the Kronecker delta function ($\delta_{jk} = 1$ if $j = k$, $\delta_{jk} = 0$ otherwise). The desired measurement residual variance can then be written by evaluating the zero-lag covariances in Eq. (E7)

$$\sigma_{\varepsilon_j}^2 = \sigma_{h^*}^2 + \sigma_{h_j}^2 \tag{E8}$$

The head variance, $\sigma_{h_j}^2$, in (E8) plays a key role in this approach since it defines how much variability one should expect around the model's predictions when the model structure and measurements are both perfect. In other words, this variance establishes a type of lower bound on the model's ability to predict point values of head (Luis and McLaughlin, 1992). If the head variance can be derived directly from the flow equation (e.g., using the solution of the statistical moment equations as presented by Zhang [1998]), Eq. (E8) can be used to evaluate the measurement residual variance to be expected when hypothesis H_0 in (E3) is true. Alternatively, one can use the numerical results of the flow model and estimate the variance of the head at each node of the discretized domain, and then use Eq. (E8) to evaluate the measurement residual variance under the assumption that H_0 is correct.

If the actual residual variance is much larger, it can be presumed that H_0 is not true (i.e., model errors are significant). Equations (E5), (E7) and (E8) suggest a few simple test statistics. One can test the assumption that the mean residual is zero (Eq. E5) and use the mean squared residual (Eq. E8) to test the null hypothesis H_0 in (E3).

E.1 Mean Residual Test

The measurement residual, ε_j , or the mismatch between observations and model predictions should have an expected value of zero at every location. If the null hypothesis is true, a sample mean computed from many measurement residuals should be close to zero. This implies a test of the following form (Luis and McLaughlin, 1992)

$$\begin{aligned} &\mathrm{H}_0 : \text{Mean residual is negligible}, \bar{\varepsilon}_j = 0 \\ &\mathrm{H}_1 : \text{Mean residual is significant}, \bar{\varepsilon}_j \neq 0 \\ &\text{Test statistic} : m_\varepsilon = \left| \frac{1}{N} \sum_{j=1}^{N} \frac{\varepsilon_j}{\sigma_{\varepsilon_j}} \right| \end{aligned} \tag{E9}$$

The decision role for this test is to decide H_0 is true if $m_\varepsilon < v$, where v is a test threshold selected to give the desired two-sided type I error probability (or significance level, α). It should be recognized that H_0 in (E9) is equivalent to H_0 in (E3). If the hypothesis is true and the measurements are sufficiently far apart for the residuals to be uncorrelated, m_ε will have a mean of zero and a standard deviation of $1/\sqrt{N}$. If we assume that m_ε is normally distributed (based

on central limit considerations), the threshold value may be readily obtained from a standard normal probability table (Luis and McLaughlin, 1992). Although the type II error is difficult to evaluate explicitly, it will decrease as *N* becomes larger, for a specified significance level (see discussion on hypothesis testing). If some of the measurements are too close for spatial correlations to be ignored (as will be the case for multiple intervals in individual boreholes), the test sample size (*N*) may be reduced to account in an approximate way for correlation effects (Luis and McLaughlin, 1992).

E.2 Mean Squared Residual Test

If one assumes that measurement residuals conform to a particular probability distribution, it would be expected that a certain percentage would lie outside confidence bounds derived from this distribution. If, for example, that distribution is normal, the interval $h_j = \hat{h}_j \pm 1.96\sigma_{\varepsilon_j}$ defines a 95% confidence interval around the predicted value $\hat{h}_j$, where σ_{ε_j} is obtained from (E8). If a significant number of the measurements h_j^* lie outside this interval, the null hypothesis H_0 is rejected. A more convenient version of the same concept relies on the following meansquared error test (Luis and McLaughlin, 1992)

$$\text{Decide } H_0 \text{ is true if: } \chi^2 = \frac{1}{N}\sum_{j=1}^{N}\frac{\varepsilon_j^2}{\sigma_{\varepsilon_j}^2} < v \qquad \text{(E10)}$$

where *v* is a test threshold selected to give the desired type I error probability (or significance level). If the hypothesis is true and the measurements are sufficiently far apart for the residuals to be uncorrelated normally distributed random variables, the test statistic χ^2 follows a chi-squared probability distribution with *N* degrees of freedom. Similar to the mean test, the type II error can be expected to decrease as *N* becomes larger, for a specified significance level. Also, the number of degrees of freedom may be reduced to account for correlation effects when the measurements are closely spaced.

E.3 Analysis of the Spatial Structure of Residuals

The statistical structure of the differences between model-predicted and observed parameters can be examined. If the examination reveals no or little correlation, the model structure is deemed acceptable, otherwise the model structure involves a systematic error (Chapman *et al.*, 1994). A series of statistical procedures can then be used to test the null hypothesis that model error is negligible.

If a significant number of the measurements are close enough to one another, it is possible to check whether or not the measurement residuals are correlated. Davis and Goodrich (1990) and Davis *et al.* (1992) propose that a model is invalid if the measurement residuals are correlated. Their criteria of acceptance are based on the change in the variance of the residuals as a function of the spatial lag or separation distance between measurement points. They use a simple semivariogram of the residuals for the analysis of variance. Using the same notations as in the previous section, this semivariogram equation can be written as

$$\gamma(l) = \frac{1}{2N(l)} \sum_{j=1}^{N(l)} \left[\varepsilon_{j+l} - \varepsilon_j \right]^2 \qquad \text{(E11)}$$

where l is the lag distance or vector, $N(l)$ is the number of data points (pairs) separated by l, and ε_{j+l} is the measurement residual at location $\mathbf{x}_j + \boldsymbol{l}$.

The analysis using the semivariogram relies on how the plot of γ changes as a function of the lag distance. If the plot is a flat horizontal line (some random variations will exist) with zero value for γ, then this is an indication of an acceptable model with perfect input and observation data. If the horizontal line has a value different than zero, it indicates an acceptable model structure, but an error in the model input, which can be adjusted or eliminated by a best-fit model prior to computing the residuals. If γ increases as a function of the lag distance, then the model is unacceptable, as there are systematic errors in the predictions. In support of this semivariogram analysis, Davis *et al.* (1992) state ."... the semivariogram analysis, while not flawless, has proved to be the most robust in terms of finding false models as invalid and true models as not invalid."

E.4 Discussion of the Stochastic Validation Approach

The three tests described in sections E.1-E.3 consider different aspects of the validation problem. The mean residual test (section E.1) checks for systematic biases (e.g., models that consistently predict much higher heads than measured). The mean-squared residual test (Section E.2) checks for overall fit (e.g., models that give head gradients and flow directions opposite to the measured). The spatial structure test (section E.3) checks for more subtle spatial features (e.g., capturing, or lack thereof, a converging flow pattern). These tests can be applied to all available measurements or to selected subsets. This gives a range of possibilities that complicates the task of reaching a conclusion about the results of a model validation. Luis and McLaughlin's (1992) view is that it is wise to examine as many performance criteria or test statistics as possible to establish an overall picture of model performance. As we mentioned earlier, we agree with this view and consider that the diversity of tests used will help evaluate different aspects of the model and establish some objective measure of the validity and confidence in the model predictions.

Although the approach outlined in this Appendix provides a quantitative measure to model validation through hypothesis testing, Luis and McLaughlin (1992) caution that this approach should not be blindly applied. In their application to the Las Cruces experiment, which has an unusually extensive set of soil data and validation measurements collected over horizontal and vertical distances of several meters and over time scales of a few years, they could not reach a conclusion regarding the ability of the model to predict the observed moisture content at later times. In addition, Ababou *et al.* (1992) assert that this approach, although very valuable, is not quite complete since the hypothesis that the model is false remains untested, and the probability of accepting a false model cannot be evaluated by this technique (Chapman *et al.*, 1994). To do this, one would need to postulate another 'complementary' model, or class of models, known to be always true if the model being tested is false. To define and implement such complementary models in an exhaustive fashion is a difficult task in the case of spatially distributed phenomena (Ababou *et al.*, 1992).

This critique of the approach of Luis and McLaughlin (1992) and of the incompleteness of hypothesis testing techniques provides more of a reason to use as many tests as possible to

evaluate model performance. As none of these statistical tests is perfect, it is beneficial to consider all these tests together and link the calibration results to the results of the validation tests for each individual realization as was shown in Figure 3. Although the possibility exists theoretically, it is highly unlikely that an individual realization that passes the majority of these tests represents a false model. If one accepts this realization as valid based on the results of these many tests, one can reasonably assume that the model user's risk, β, is very small. On the other hand, if an individual realization fails to pass a large number of these tests, then rejecting this realization for being invalid is not expected to represent a large type I error.

Appendix F: Sequential Self-Calibration (SSC) Approach

To continue reducing the uncertainty level, a refinement of the conductivity distribution can be made using the SSC method. In this method, new head measurements (and old ones) can be used to condition the generation of the conductivity field in such a way that the uncertainty in the conductivity heterogeneity pattern around each measurement location is reduced.

Particular interest arises for conditioning on the head and concentration data in the numerical analyses due to the fact that these data carry important information on the spatial variation and, more importantly, the spatial hydraulic connectivity (flow channels or barriers) that may not be captured by traditional hydraulic conductivity data. Several new geostatistically-based inverse approaches have been developed to generate the hydraulic conductivity fields by conditioning on both the hydraulic head and conductivity measurements (Zimmerman *et al.*, 1998). Among them, the SSC method (Sahuquillo *et al.*, 1992; Gómez-Hernández *et al.*, 1997; Capilla *et al.*, 1997, 1998; Wen *et al.*, 1996, 1999) is an iterative geostatistically-based inverse technique that allows generation of multiple equiprobable realizations of heterogeneous fields that match the dynamic data, in addition to the typical geostatistical constraints. It has been demonstrated to be computationally efficient for modeling hydraulic conductivity field and capable of identifying flow channels embedded in the aquifer by conditioning on multiple production well data (Wen *et al.*, 1999). Current SSC method is developed for the analysis of reservoir permeability of oil fields conditioning on oil and water flow rates. In the validation process, we can use this method in the refinement portion of the iterative loop of modeling-validation-refinement. This method, or any other similar one, can be systematically used to help the dual purpose of refining the model predictions and reducing their uncertainty bounds.

The main steps in the SSC method are adapted and summarized here within the application to the validation approach. First, one would start with the original hydraulic conductivity fields generated with the original model that is yet to be evaluated and validated. Using the flow and transport solutions provided by the original model for each individual realization, one would process these realizations one at a time utilizing the new (as well as old) collected data for validation purposes. An objective function that measures the mismatch between predicted and observed head and concentration data can then be written as (e.g., Wen *et al.*, 1999)

$$O = \sum_{1}^{nwell} W_c(nw)[\hat{C}(nw) - C(nw)]^2 + \sum_{1}^{nwell} W_h(nw)[\hat{h}(nw) - h(nw)]^2 \qquad \text{(F1)}$$

where $W_c(nw), W_h(nw)$ are the weights assigned to the concentration and head sampling well nw according to sampling accuracy. Matching the head and concentration data is achieved by minimizing this objective function. A gradient-based method is used for optimization, which requires the calculation of "sensitivity coefficients," the derivatives of the concentration and head with respect to the hydraulic conductivity perturbation:

$$\frac{\partial \hat{C}(nw)}{\partial \Delta K_i}, \frac{\partial \hat{h}(nw)}{\partial \Delta K_i} \quad i = 1, \cdots, N \qquad \text{(F2)}$$

where N is the number of blocks in the model. In practice, the number of actual blocks within which conductivity is perturbed can be reduced to between 1/100 and 1/10 of the number of blocks of the entire domain using the "master point" concept (Gómez-Hernández *et al.*, 1997).

The optimal changes of conductivity are determined at these master points and then smoothly interpolated by kriging to all grid blocks. The sensitivity coefficients are calculated as part of the solutions of the flow and transport equations by book-marking each particle's trajectory and travel times. The next step is to determine the optimal perturbations of the conductivity at all master locations with a gradient projection-based method. The optimal conductivity perturbations at the master locations are then smoothly propagated to all grid cells by kriging. One would then go back and evaluate the objective function until it is sufficiently close to zero, or less than a predetermined tolerance value. Fewer than 20 iterations are normally required (Wen *et al.*, 1999).

The unique feature of the SSC method is that it results in multiple equiprobable realizations of hydraulic conductivity fields that match observed head and concentration data and are consistent with the spatial statistics of the initial conductivity field realizations. This will help reduce the uncertainty bound on model predictions (by reducing conductivity distribution uncertainty) and guide convergence of the realizations reduction process to a very compact and representative set. In addition, fast computation of sensitivity coefficients within one single simulation makes inversion feasible.